# FedRLHF: A Convergence-Guaranteed Federated Framework for Privacy-Preserving and Personalized RLHF

Flint Xiaofeng Fan
National University of Singapore
Singapore
fxf@u.nus.edu

Cheston Tan
CFAR, A*STAR
Singapore
cheston-tan@i2r.a-star.edu.sg

Yew-Soon Ong
CFAR, A*STAR
Singapore
asysong@ntu.edu.sg

Roger Wattenhofer
ETH Zurich
Zurich, Switzerland
wattenhofer@ethz.ch

Wei-Tsang Ooi
National University of Singapore
Singapore
ooiwt@comp.nus.edu.sg

## ABSTRACT

In the era of increasing privacy concerns and demand for personalized experiences, traditional Reinforcement Learning with Human Feedback (RLHF) frameworks face significant challenges due to their reliance on centralized data. We introduce Federated Reinforcement Learning with Human Feedback (FedRLHF), a novel framework that decentralizes the RLHF process. FedRLHF enables collaborative policy learning across multiple clients, such as Large Language Models (LLMs) finetuning, without sharing raw data or human feedback, thereby ensuring robust privacy preservation. Leveraging federated reinforcement learning, each client integrates human feedback locally into reward functions and updates their policies through personalized RLHF processes. We establish rigorous theoretical foundations for FedRLHF, providing convergence guarantees, and deriving sample complexity bounds that scale efficiently with the number of clients. Empirical evaluations on the MovieLens and IMDb datasets demonstrate that FedRLHF preserves user privacy, achieves performance on par with centralized RLHF, and enhances personalization across diverse client environments.

## KEYWORDS

Federated RL; RLHF; LLMs; Personalization; Privacy-preserving AI

**ACM Reference Format:**
Flint Xiaofeng Fan, Cheston Tan, Yew-Soon Ong, Roger Wattenhofer, and Wei-Tsang Ooi. 2025. FedRLHF: A Convergence-Guaranteed Federated Framework for Privacy-Preserving and Personalized RLHF. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 26 pages.

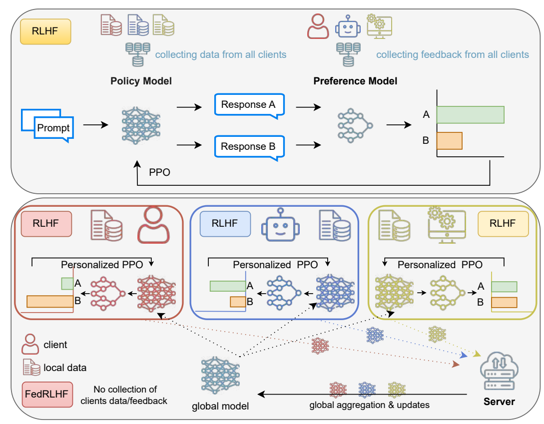



**Figure 1: Comparison of the FedRLHF framework to conventional RLHF methods. Top: Conventional RLHF requires centralized collection of user data and feedback to train the policy model and preference (reward) model. Bottom: In FedRLHF, clients maintain local policy models trained on-device using RLHF with local data and preference models. Only policy model updates are shared with a central server, which aggregates them to refine a global policy.**

## 1 INTRODUCTION

Reinforcement Learning with Human Feedback (RLHF) has emerged as a powerful paradigm for training intelligent agents that align closely with human values and preferences [3, 26]. By integrating human feedback into the reinforcement learning loop, RLHF has enabled significant advancements in natural language processing, robotics, and personalized recommendation systems [4, 37, 50]. A prominent example is ChatGPT [25], where RLHF has been instrumental in fine-tuning large language models (LLMs) to generate more coherent, contextually appropriate, and user-aligned responses [49].

Despite these successes, the practice of aggregating data and feedback from multiple users in centralized RLHF systems poses significant *privacy* risks, especially in domains involving sensitive information such as healthcare or finance [35]. For instance, a personalized healthcare assistant using RLHF requires centralizing patient data and feedback, potentially violating regulations like Health Insurance Portability and Accountability Act (HIPAA) [38] and General Data Protection Regulation (GDPR) [8], while exposing users to risks of data breaches and identity theft. Moreover, different organizations or individuals may be reluctant to share their feedback data due to intellectual property concerns or competitive advantages [19].



*Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA*. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

In addition to privacy risks, centralization creates substantial hurdles for achieving *personalization* in RLHF systems. Users exhibit diverse preferences and behaviors, making it challenging for a centralized policy to cater effectively to all individuals [3]. Balancing global performance with personalization becomes non-trivial, as optimizing for the average user can lead to suboptimal experiences for specific individuals [36]. Returning to our healthcare assistant example, patients may have unique health conditions and treatment preferences. A one-size-fits-all model, trained on centralized data, may fail to provide the personalized recommendations necessary for optimal care, potentially impacting patient outcomes negatively.

To address these limitations, we propose **Federated Reinforcement Learning with Human Feedback (FedRLHF)**, a novel federated framework that uniquely integrates RLHF principles with federated reinforcement learning [5, 9] to simultaneously address privacy and personalization challenges. As illustrated in Figure 1, FedRLHF decentralizes the RLHF process so that each client updates a local policy using only its own data and private human feedback. By exchanging only model updates rather than raw data, FedRLHF preserves privacy and adheres more closely to data governance regulations. Moreover, each client can personalize its learning process by shaping the local reward function with individual human feedback, ensuring that policies align with user preferences. This decentralized approach supports collaboration across clients while maintaining privacy and personalization.

The FedRLHF framework introduces unique technical challenges, notably ensuring *convergence guarantees* in a decentralized environment and managing the *trade-off between global performance and personalization*. In federated reinforcement learning, the variability in clients' environments and behaviors can lead to instability and divergence in the learning process. Our FedRLHF framework addresses these challenges in a principled manner. We provide a comprehensive theoretical analysis establishing convergence guarantees and deriving bounds on the sample complexity under standard assumptions. Furthermore, we introduce a quantitative measure of personalization and a personalization parameter in the reward shaping function to analyze and control the trade-off between global policy alignment and personalized adaptation.

Our contributions are as follows:

- We introduce FedRLHF, a framework that integrates federated reinforcement learning with human feedback, enabling privacy-preserving and personalized policy learning and model fine-tuning (Section 3).
- We provide convergence guarantees and derive sample complexity bounds that account for the integration of human feedback, extending existing FedRL theory (Section 4).
- We develop a quantitative measure of personalization to analyze the trade-off between maximizing global performance and adapting individual client policies (Section 5).
- We empirically demonstrate FedRLHF's effectiveness on the MovieLens and IMDb datasets, showcasing its ability to preserve privacy, match centralized RLHF performance, and enhance personalization (Section 6).
- A full version of this work, including additional proofs and experimental details, is available at [11], and our code is publicly available at github.com/flint-xf-fan/Federated-RLHF.

## 2 BACKGROUND & RELATED WORK

**Reinforcement Learning with Human Feedback** (RLHF) has become instrumental in aligning machine learning models with human values and preferences [3, 26, 37, 49]. By integrating human-generated feedback into the reward structure, RLHF facilitates the training of policies that exhibit behaviors more aligned with human preferences. However, traditional RLHF frameworks predominantly rely on centralized aggregation of data and feedback, which introduces significant privacy concerns. In addition, the centralization nature of RLHF also limits personalization, as it typically involves training a single reward model for all clients. This approach fails to accommodate the heterogeneous preferences of individual clients, leading to suboptimal policy performance in diverse environments, as illustrated in Figure 1.

While previous studies [28] have explored personalized reward models in RLHF through representation learning, they still depend on the centralized aggregation of user data and feedback, failing to address the privacy concerns inherent in centralized systems. Additionally, methods like Group Robust Preference Optimization (GRPO) [31] aim to reduce performance disparities across user groups but similarly rely on aggregated, centralized datasets and do not provide mechanisms for personalization at the individual level. A recent effort by Li et al. [20] introduces a framework for personalized language modeling from personalized human feedback, which learns user-specific embeddings to capture individual preferences. However, their approach still relies on centralized data collection and does not address privacy concerns. Moreover, these prior approaches do not simultaneously tackle both privacy preservation and personalization challenges in real-world scenarios where user data is distributed across multiple devices or organizations.

**Federated Reinforcement Learning** (FedRL) has garnered significant attention in recent years, aiming to leverage the principles of federated learning [24] across diverse RL clients to enhance their sample efficiency without sharing raw data or trajectories of the sequential decision-making process. This approach has shown promise in diverse applications, from optimizing autonomous vehicles and enhancing edge caching in IoT networks to smart management of building facibilities [6, 12, 21, 23, 41, 47, etc.]. Recent theoretical advancements have solidified the foundations of FedRL. Notably, convergence guarantees and sample complexity bounds have been established, demonstrating speedup with increasing numbers of participating agents [9]. In addition, the application of Markovian sampling techniques has been shown to achieve linear convergence speedup in FedRL settings [17]. Furthermore, recent analysis of decentralized FedRL with Byzantine fault tolerance has proven fast convergence rates without relying on a central server, marking a significant step towards fully distributed and resilient RL systems [15, 29]. Recent works have explored the benefits of heterogeneity among Q-learning agents [10]. Woo et al. [44] prove that leveraging agent heterogeneity can lead to linear speedup and significant efficiency gains in federated Q-learning settings. In addition, Woo et al. [45] introduce a federated offline RL method that achieves linear speedup with low communication costs in heterogeneous client environments. Moreover, Wang et al. [40] leverage momentum mechanisms to achieve exact convergence and state-of-the-art sample efficiency in highly heterogeneous environments.

Finally, Jiang et al. [14] propose the first policy distillation-based framework that aligns heterogeneous agent policies to further accelerate convergence and enhance sample efficiency in federated policy gradient methods.

**Novelty of Our Approach.** In contrast to existing FedRL methods that employ a uniform reward structure, our approach, FedRLHF, integrates client-specific human feedback directly into the reward function. By allowing each client to locally shape its reward, FedRLHF not only preserves privacy (since sensitive feedback remains on-device), but also enhances personalization by accommodating diverse user preferences. This personalized feedback mechanism is particularly transformative for LLM applications (e.g., ChatGPT), as it enables context-aware fine-tuning of responses while upholding stringent privacy standards. Moreover, our framework is theoretically distinguished by convergence guarantees and sample complexity bounds that explicitly capture the variability introduced by human feedback—an aspect absent from prior FedRL literature.

## 3 PROBLEM FORMULATION

We consider a federated reinforcement learning system with $K$ clients, where each client $k \in 1, 2, \ldots, K$ interacts with its own environment, modeled as a Markov Decision Process (MDP) $M_k = (\mathcal{S}, \mathcal{A}, P_k, R_k, \rho_0(s), \gamma)$. Here, $\mathcal{S}$ and $\mathcal{A}$ represent the state and action spaces, respectively. The state-transition function $P_k : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ defines the probability $P_k(s' \mid s, a)$ of transitioning from state $s$ to state $s'$ after taking action $a$. The reward function $R_k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ specifies the expected reward $R_k(s, a)$ for client $k$ when action $a$ is taken in state $s$. Finally, $\gamma \in [0, 1)$ is the discount factor balancing immediate and future rewards. The initial state distribution $\rho_0(s)$ specifies the probability of the MDP starting in state $s$. We assume that both $\gamma$ and $\rho$ are fixed and known for all clients. This uniformity simplifies the theoretical analysis and ensures consistency in policy evaluation and optimization. However, extensions to client-specific discount factors and initial state distributions are straightforward within our framework.

Each client's MDP may vary in the transition dynamics $P_k$ and reward functions $R_k$, reflecting the heterogeneity among clients due to personalized environments or preferences. Let $\pi_\theta : \mathcal{S} \to \Delta(\mathcal{A})$ denote a stochastic policy parameterized by $\theta \in \mathbb{R}^d$, where $\Delta(\mathcal{A})$ is the set of probability distributions over the action space $\mathcal{A}$. The policy $\pi_\theta(a \mid s)$ specifies the probability of taking action $a$ in state $s$ under the parameters $\theta$. For each client $k$, the objective is to find the policy parameters $\theta$ that maximize the expected cumulative discounted reward:

$$J_k(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t R_k(s_t, a_t)\right], \tag{1}$$

where the expectation is taken over trajectories $\tau = (s_0, a_0, s_1, \ldots)$ generated by following policy $\pi_\theta$ in $M_k$ , starting from $\rho_o(s)$.

### 3.1 Incorporating Human Feedback

Unlike conventional RLHF where a single reward function and policy are learned from aggregated data, FedRLHF allows for client-specific reward functions $R_k$ and locally adapted policies $\pi_k$. In FedRLHF, human feedback is integrated locally to shape each client's reward function. Specifically, the reward function for client $k$ is augmented as:

$$R_k(s, a) = R_k^0(s, a) + \lambda H_k(s, a), \tag{2}$$

where $R_k^0(s, a)$ is the intrinsic reward provided by the environment, $H_k(s, a)$ is the client-specific human feedback function representing additional reward or penalty based on human evaluation, and $\lambda > 0$ is a scaling factor balancing the influence of human feedback relative to the intrinsic reward.

**Algorithm 1** FedRLHF

**Require:** Number of clients $K$, total communication rounds $T$, local update steps $\tau$, personalization factor $\lambda$, learning rate $\eta$

**Ensure:** Final global policy parameters $\theta_{\text{final}}$

1: Initialize global policy parameters $\theta_0$
2: **for** $t = 0$ **to** $T - 1$ **do**
3:     **Server** broadcasts global parameters $\theta_t$ to all clients
4:     **for each client** $k \in \{1, 2, \ldots, K\}$ **in parallel do**
5:         Initialize local parameters: $\theta_{t,0}^k \leftarrow \theta_t$
6:         **for** $i = 0$ **to** $\tau - 1$ **do**
7:             Sample a mini-batch $\mathcal{B}_{t,i}^k$ using policy $\pi_{\theta_{t,i}^k}$ in $M_k$
8:             Collect human feedback $H_k(\mathcal{B}_{t,i}^k)$
9:             Calculate shaped reward $R_k$ per Eq. (2)
10:             Estimate policy gradient per $J_k(\theta_{t,i}^k)$ in Eq. (1):
$$\hat{g}_{t,i}^k \leftarrow \nabla_\theta \hat{J}_k(\theta_{t,i}^k; \mathcal{B}_{t,i}^k)$$
11:             Update local parameters:
$$\theta_{t,i+1}^k \leftarrow \theta_{t,i}^k + \eta \hat{g}_{t,i}^k$$
12:         Compute local model update:
$$\Delta\theta_{t+1}^k \leftarrow \theta_{t,\tau}^k - \theta_t$$
13:     **Server** aggregates local updates:
$$\theta_{t+1} \leftarrow \theta_t + \frac{1}{K}\sum_{k=1}^{K} \Delta\theta_{t+1}^k$$
14: **return** $\theta_{\text{final}} \leftarrow \theta_T$ or $\frac{1}{T}\sum_{t=0}^{T-1} \theta_t$

### 3.2 Global Objective

The global objective is to find policy parameters $\theta$ that maximize the average expected cumulative reward across all clients:

$$J(\theta) = \frac{1}{K}\sum_{k=1}^{K} J_k(\theta). \tag{3}$$

### 3.3 The FedRLHF Framework

We propose the **Federated Reinforcement Learning with Human Feedback (FedRLHF)** framework to optimize the global objective (3) across all clients in a federated manner while respecting their individual environments and preferences. Algorithm 1 presents the pseudocode for FedRLHF. The key components and operations of the framework are as follows:

*Local RLHF (lines 6-12).* : The FedRLHF framework allows clients to use different RL methods, including Q-learning [42] and policy gradient (PG) [43], to perform $\tau$ steps of local optimization. The theoretical analysis provided in this manuscript assumes PG methods for local updates, which was necessary to facilitate the convergence analysis and derive sample complexity bounds.

*Reward Shaping (line 9).* : Rewards are shaped as $R_k = R_k^0 + \lambda H_k$, where $R_k^0$ is the intrinsic reward and $\lambda$ controls the influence of human feedback $H_k$.

*Privacy Preservation.* : Only the model updates $\Delta\theta_{t+1}^k$ are transmitted to the server, preventing direct access to individual user data and feedback, thus providing a significant level of privacy protection. For applications requiring formal privacy guarantees or protection against advanced inference attacks, additional measures such as Differential Privacy (DP) [7] could be incorporated into the framework.

*Aggregation (line 13).* : Multiple server aggregation methods exist, such as FedAvg (simple averaging) [24], Weighted Average based on client data sizes, and robust aggregation techniques like median-based methods [46]. The choice depends on specific requirements such as fairness, robustness to outliers, or heterogeneity in client data. In this algorithm and our theoretical analysis, we employ FedAvg for its simplicity and to facilitate clearer theoretical results. However, our framework is flexible and can accommodate other aggregation methods if needed.

*Mechanism of Personalization.* Although the server broadcasts a single global policy parameter vector, personalization arises because each client's local objective includes its own reward function, shaped by both intrinsic rewards and the local human feedback function $H_k$. At each round, client $k$ adapts the global parameters to better optimize $\lambda H_k$ alongside $R_k^0$. This leads to client-specific parameter updates. While the global model aggregates these updates to improve overall performance, each client's environment and feedback distribution remain unique. In practice, clients can also maintain partially fine-tuned local parameters or personalized embeddings, thus capturing their distinct preferences or constraints. Consequently, even though the server aggregates all updates, the local RLHF step ensures that each client's policy is guided by its private human feedback, achieving personalization in a federated manner.

# 4 CONVERGENCE RESULTS

In this section, we provide theoretical guarantees on the convergence and sample complexity of the FedRLHF framework. These results apply to implementations of the framework that adhere to the core principles and structure outlined in Algorithm 1, under the necessary assumptions stated below:

## 4.1 Assumptions

ASSUMPTION 1 ($L$-SMOOTH GRADIENTS). *For all $\theta, \theta' \in \mathbb{R}^d$ and $k \in [K]$, the gradients of the clients' objective functions are $L$-Lipschitz continuous:*

$$\|\nabla J_k(\theta) - \nabla J_k(\theta')\| \leq L\|\theta - \theta'\|.$$

ASSUMPTION 2 ($G$-BOUNDED GRADIENTS). *For all $\theta \in \mathbb{R}^d$ and $k \in [K]$, the gradients are bounded:*

$$\|\nabla J_k(\theta)\| \leq G.$$

ASSUMPTION 3 ($\sigma$-BOUNDED VARIANCE). *For all $\theta \in \mathbb{R}^d$ and $k \in [K]$, the variance of the stochastic gradient estimator is bounded:*

$$\mathbb{E}\left[\|\nabla \hat{J}_k(\theta) - \nabla J_k(\theta)\|^2\right] \leq \sigma^2,$$

*where $\nabla \hat{J}_k(\theta)$ is the stochastic gradient computed from a mini-batch.*

ASSUMPTION 4 (BOUNDED SECOND MOMENT). *For all $\theta \in \mathbb{R}^d$ and $k \in [K]$, the second moment of the stochastic gradient is bounded:*

$$\mathbb{E}\left[\|\nabla \hat{J}_k(\theta)\|^2\right] \leq M^2.$$

ASSUMPTION 5 (POLYAK-ŁOJASIEWICZ (PL) CONDITION). *The global objective function satisfies the PL condition:*

$$2\mu\left(J(\theta^*) - J(\theta)\right) \leq \|\nabla J(\theta)\|^2, \quad \forall\theta \in \mathbb{R}^d,$$

*where $\mu > 0$ is a constant and $\theta^* = \arg\max_\theta J(\theta)$.*

REMARK. *Assumptions 1–4 are common in the stochastic optimization literature. The PL condition (Assumption 5) is stronger, especially for reinforcement learning's typically non-convex objectives. However, it approximates scenarios where objective functions exhibit properties conducive to linear convergence. Policy gradient methods with trust region constraints, such as TRPO [33], or those using proximal objectives, like PPO [34], often result in smoother updates to the policy parameters, making the PL condition more reasonable. Recent works [2, 16, 48, etc.] have used the PL condition for non-convex convergence guarantees for RL, further justifying its use in our analysis.*

ASSUMPTION 6 (BOUNDED HUMAN FEEDBACK). *For all $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $k \in [K]$, the human feedback is bounded:*

$$|H_k(s, a)| \leq H_{\max}.$$

REMARK. *Assumption 6 limits the variance introduced by human feedback in the learning process. In our experiments with the MovieLens task, we implement this by bounding feedback values and options (Section 6.1.2), similar to practical systems like ChatGPT that curate feedback for consistency.*

## 4.2 Convergence and Sample Complexity

We now present the main theoretical results, starting with key lemmas leading up to the convergence theorem. The complete proof for theorems presented in this section is provided in Appendix B.

LEMMA 4.1 (BOUNDED LOCAL-GLOBAL DIFFERENCE). *Under Assumptions 1, 2, and 3, for any communication round $t$ and client $k$, we have:*

$$\mathbb{E}\left[\|\theta_t^k - \theta_t\|^2\right] \leq \eta^2\tau^2(G^2 + \sigma^2)$$

*where $\theta_t^k$ is the local model of client $k$, $\theta_t$ is the global model, $\eta$ is the learning rate, $\tau$ is the number of local updates, $G$ is the gradient bound, and $\sigma^2$ is the variance bound.*

REMARK. *Lemma 4.1 quantifies the extent to which local models diverge from the global model after $\tau$ local updates. This deviation is influenced by the learning rate $\eta$ and the number of local updates $\tau$, both of which amplify the divergence when increased.*

Lemma 4.2 (One-Step Descent). *Under Assumptions 1–6, for any round $t$, the expected improvement in the global objective satisfies:*

$$\mathbb{E}[J(\theta_{t+1})] \geq J(\theta_t) + \eta\tau\left(1 - \frac{L\eta\tau}{2}\right)\|\nabla J(\theta_t)\|^2 - \frac{L}{2}\left(\frac{\eta^2\tau^2}{K}\right)(G^2+\sigma^2) - \lambda H_{\max}$$

*where $\theta_{t+1}$ is the updated global model.*

Remark. *This lemma establishes that each communication round in FedRLHF yields a quantifiable improvement in the global objective $J(\theta)$. The positive term $\eta\tau\left(1-\frac{L\eta\tau}{2}\right)\|\nabla J(\theta_t)\|^2$ signifies progress towards maximizing the objective, while the negative terms $\lambda H_{\max}$ and $\frac{L}{2}\left(\frac{\eta^2\tau^2}{K}\right)(G^2+\sigma^2)$ account for the inherent variance in stochastic gradients and the bounded impact of human feedback, respectively.*

Theorem 4.1 (Convergence of FedRLHF). *Under Assumptions 1–6, if we choose the constant learning rate $\eta = \frac{1}{L\tau}$, then the output $\theta_{avg} = \frac{1}{T}\sum_{t=0}^{T-1}\theta_t$ of Algorithm 1 satisfies:*

$$\mathbb{E}\left[J(\theta^*) - J(\theta_{avg})\right] \leq \frac{L}{\mu T}\left(J(\theta^*) - J(\theta_0)\right) + \frac{1}{2\mu K}(G^2+\sigma^2) + \frac{L}{\mu}\lambda H_{\max}.$$

Theorem 4.1 establishes that the FedRLHF algorithm converges to an optimal policy within a bounded suboptimality gap. This bound elucidates several key aspects of the algorithm's performance:

(1) **Convergence Rate ($O(1/T)$):** The first term $\frac{L}{\mu T}\left(J(\theta^*) - J(\theta_0)\right)$ indicates that the algorithm achieves a linear convergence rate with respect to the number of communication rounds $T$, which aligns well with expectations in federated optimization under the PL condition.
(2) **Impact of Client Diversity and Variance ($O(1/K)$):** The second term $\frac{1}{2\mu K}(G^2+\sigma^2)$ scales inversely with the number of clients $K$. This indicates that aggregating updates from more clients reduces the effect of gradient variance $\sigma^2$ and bounded gradient norms $G$, leading to a tighter convergence bound.
(3) **Influence of Human Feedback ($O(1)$):** The third term $\frac{L}{\mu}\lambda H_{\max}$ represents the bounded influence of human feedback on the convergence. The scaling factor $\lambda$ and the maximum human feedback bound $H_{\max}$ determine how much human feedback can potentially offset the objective's improvement. While human feedback guides the policy towards user preferences, this term ensures that its impact remains controlled, preventing excessive deviations that could hinder convergence.

Theorem 4.2 (Sample Complexity of FedRLHF). *Under Assumptions 1–6, to achieve an expected optimality gap of*

$$\mathbb{E}\left[J(\theta^*) - J(\theta_{avg})\right] \leq \epsilon,$$

*the total number of samples required across all clients is:*

$$N = O\left(\frac{L(G^2+\sigma^2)}{\mu^2\epsilon^2}\right),$$

*subject to:*

$$K \geq O\left(\frac{G^2+\sigma^2}{\mu\epsilon}\right), \qquad \lambda H_{\max} \leq O\left(\frac{\mu\epsilon}{L}\right).$$

Theorem 4.2 provides an estimate of the total number of samples required across all clients to achieve a desired expected optimality gap $\epsilon$. The bound, which scales with $O(\epsilon^{-2})$ and aligns with standard results in stochastic optimization, reveals several insights:

(1) **Dependence on Problem Constants:** The constants $\mu$ and $L$ reflect the curvature properties of the objective function $J(\theta)$. A larger $\mu$ (stronger PL condition) and smaller $L$ (less curvature) lead to a lower sample complexity. The term $G^2+\sigma^2$ captures the combined effect of the gradient bound and the variance of the stochastic gradients. Reducing them through techniques like gradient clipping or variance reduction methods [27] can significantly lower sample complexity.
(2) **Per-Client Sample Complexity:** Diving the bound by $K$ yields the sample complexity per client: $N_c = \frac{N}{K} = O\left(\frac{L(G^2+\sigma^2)}{K\mu^2\epsilon^2}\right)$. As $K$ increases, the per-client sample complexity $N_c$ decreases proportionally, suggesting that individual clients require fewer samples to achieve the same level of accuracy when more clients participate. This per-client sample efficiency gain aligns with the collaborative nature of federated reinforcement learing [9].
(3) **Cost of Personalization:** As detailed in the proof in Appendix B, a larger $\lambda H_{\max}$ increases $\epsilon_H$, effectively consuming more of the allowable error budget. If $\epsilon_H$ becomes significant relative to $\epsilon$, the remaining error budget for $\epsilon_T$ and $\epsilon_V$ decreases. This necessitates tighter convergence in these terms, requiring a larger number of communication rounds $T$ and potentially more clients $K$. Consequently, the total sample complexity $N$ increases indirectly due to a higher emphasis on personalization, suggesting a trade-off between personalization and efficiency

# 5 PERSONALIZATION-PERFORMANCE TRADE-OFF ANALYSIS

In this section, we establish a formal relationship between personalization and global performance within the FedRLHF framework. The convergence analysis in Section 4 already hints at the personalization-performance trade-off, particularly through the influence of the human feedback weight $\lambda$ on the convergence. Here, we provide a quantitative measure of this trade-off by analyzing how personalization affects global performance of intrinsic rewards and sample complexity. The complete proof for theorems presented in this section is provided in Appendix C.

## 5.1 Definitions and Preliminaries

Definition 5.1 (Maximum Reward). *We define $R_{max}$ as the maximum absolute value of the intrinsic reward function across all clients and state-action pairs:*

$$R_{max} = \max_{k\in\{1,2,\ldots,K\},\, s\in\mathcal{S},\, a\in\mathcal{A}} |R_k^0(s,a)|.$$

Definition 5.2 (Personalization Score). *For a client $k$ with policy $\pi_k(\cdot|s,\theta)$ and the global policy $\pi(\cdot|s,\theta)$, the personalization score is defined as:*

$$P_k(\theta) = \mathbb{E}_{s\sim\rho}\left[D_{KL}\left(\pi_k(\cdot|s,\theta) \parallel \pi(\cdot|s,\theta)\right)\right],$$

*where $\rho$ is the state distribution and $D_{KL}$ denotes the Kullback-Leibler divergence.*

Definition 5.3 (Global Performance Metric). *We define the global performance metric as the average expected cumulative intrinsic reward across all clients:*

$$J_g(\theta) = \frac{1}{K}\sum_{k=1}^{K} J_k^0(\pi),$$

*where*

$$J_k^0(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t R_k^0(s_t, a_t)\right],$$

*and the expectation is over trajectories $\tau = (s_0, a_0, s_1, \ldots)$ generated by following policy $\pi$ in client $k$'s MDP $M_k$, starting from $\rho_0(s)$.*

## 5.2 Personalization-Performance Trade-off

We now present our main theorem on the trade-off between personalization and global performance.

Theorem 5.1 (Personalization-Performance Trade-off). *Under Assumptions 1–6 and Definition 5.1–5.3, for any set of client policies $\{\pi_k(\cdot|s,\theta)\}_{k=1}^K$ and the global policy $\pi(\cdot|s,\theta)$, the global performance metric satisfies:*

$$J_g(\theta) \geq \frac{1}{K}\sum_{k=1}^{K} J_k^0(\pi_k) - C\cdot\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{P_k(\theta)}\right),$$

*where $C > 0$ is a constant given by: $C = \frac{2\sqrt{2}R_{total,\max}}{(1-\gamma)^2}$, and $R_{total,\max} = R_{max} + \lambda H_{\max}$ is the maximum possible total reward.*

Theorem 5.1 establishes that the global performance $J_g(\theta)$ is lower bounded by the average client-specific performance (intrinsic rewards) $\frac{1}{K}\sum_{k=1}^K J_k^0(\pi_k)$ minus a penalty term proportional to the average of the square roots of the personalization scores $\frac{1}{K}\sum_{k=1}^K \sqrt{P_k(\theta)}$. The constant $C$ encapsulates the maximum possible total reward and the discount factor, indicating that in environments with higher rewards or longer planning horizons, the impact of personalization on global performance is more pronounced.

## 5.3 Impact of Human Feedback

We analyze how the incorporation of human feedback, governed by the weight $\lambda$, influences personalization and global performance.

Theorem 5.2 (Impact of Human Feedback). *Under the same assumptions and definitions in Theorem 5.1, as the human feedback weight $\lambda$ increases:*

1. *The average personalization score $\frac{1}{K}\sum_{k=1}^K P_k(\theta)$ increases at a rate of $O(\lambda^2)$.*
2. *The global performance $J_g(\theta)$ decreases at a rate of $O(\lambda)$.*
3. *The sample complexity $N$ increases at a rate of $O(\lambda)$.*

Theorem 5.2 quantitatively demonstrates that increasing the human feedback weight $\lambda$ intensifies personalization (as the personalization score grows at $O(\lambda^2)$) but leads to a linear decrease in global performance and an increase in sample complexity.

1. **Personalization Score Increases at $O(\lambda^2)$:** The personalization score $P_k(\theta)$ for each client scales quadratically with $\lambda$, indicating that the degree of personalization becomes more pronounced as $\lambda$ increases.
2. **Global Performance Decreases at $O(\lambda)$:** The global performance $J_g(\theta)$ experiences a linear decrease with respect to $\lambda$. This implies that while personalization enhances client-specific performance, it concurrently introduces a controlled degradation in overall system performance.
3. **Sample Complexity Increases at $O(\lambda)$:** The total number of samples $N$ required to achieve a desired level of performance grows linearly with $\lambda$. This reflects the increased data demands associated with higher levels of personalization to maintain convergence guarantees.

# 6 EMPIRICAL RESULTS

We evaluate FedRLHF's effectiveness in integrating human feedback within a federated reinforcement learning setting through two real-world tasks: movie rating prediction using the MovieLens dataset and sentiment-controlled review generation using the IMDb dataset. Our experiments benchmark FedRLHF against a centralized RLHF baseline, with a focus on personalization, and maintaining performance levels.

All experiments were conducted on an NVIDIA GeForce RTX 3090 GPU, using the Flower framework [1] to simulate a realistic federated learning environment with gRPC communication, mimicking real-world distributed systems. Detailed experimental results and analyses are provided in Appendix D.

## 6.1 Movie Rating Prediction on MovieLens

*6.1.1 Task Description and Setup.* In this task, we simulate a streaming service enhancing its recommendation system while preserving user privacy and catering to individual preferences. Using the `ml-latest-small` version of the MovieLens dataset [13] which contains 100,836 ratings from 610 users on 9,742 movies, we randomly selected $K = 10$ users as clients, each with unique viewing histories and preferences. The objective is to predict whether a user would assign a high rating (4 stars or above) to a given movie, effectively framing this as a binary classification task.

*6.1.2 Human Feedback Simulation.* To emulate realistic user behavior and feedback mechanisms, we developed a noise-aware, rule-based feedback simulator generating two types of feedback: **a. Direct Feedback:** Categorizes predictions as "too high" (-1), "too low" (1), or "about right" (0) based on the difference between predicted and actual ratings; **b. Comparative Feedback:** Expresses preferences between movie pairs, mirroring real-world scenarios where users more easily compare options than provide absolute ratings. Feedback values are bounded within [-1, 1], satisfying Assumption 6. This feedback trains a local reward (preference) model for each client. Full details on the feedback simulation are provided in Appendix D.1.2.

*6.1.3 Implementation.* We implemented a neural network model with embedding layers for users and movies. The model inputs included user IDs, movie IDs, and movie genre information to capture complex user-movie interactions. In the federated learning process, each client trained the model locally using intrinsic rewards and simulated human feedback, employing Q-learning as the local RLHF step. We employ Q-learning in this task due to its effectiveness in handling discrete action spaces (movie recommendations) and its

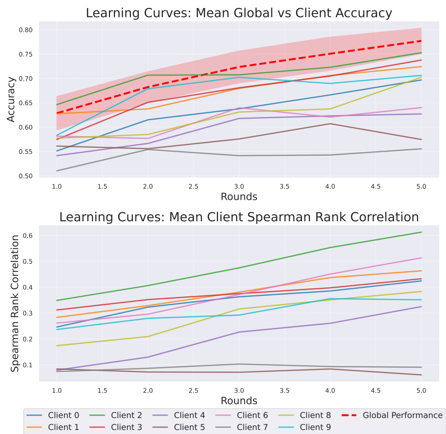



**Figure 2: Learning curves on MovieLens: (top) Global vs. Client Accuracy, (bottom) Client Spearman correlation.**

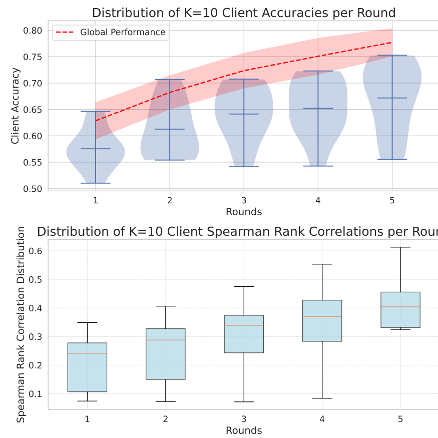



**Figure 3: Distribution of $K = 10$ client accuracies and Spearman rank correlations per round for the MovieLens task.**

ability to learn optimal policies in environments with delayed rewards. Clients performed 5 local epochs per federated round, using the Adam optimizer [18] (learning rate $1 \times 10^{-3}$). Global model aggregation used a weighted average based on client example counts, following a FedAvg variant [24]. The process spanned 5 communication rounds. More details are provided in Appendix D.1.3.

*6.1.4 Results and Analysis.* Figure 2 presents the learning curves for both global and client-specific accuracies (top), along with the Spearman rank correlations (bottom) for each of the $K = 10$ clients across 5 federated rounds. All results are averaged over five independent runs. For clarity, only client means across runs are shown. Global performance is depicted by the red dashed line (mean) with shaded area (standard deviation).

*Global Performance Improvements.* The global performance in accuracy improves from 62.86% ± 3.45% to 77.71% ± 2.64% over 5 rounds. The steady improvement in global performance, which is also evident from the client-specific accuracies distribution shown in the violin plot in Figure 3 (top subplot), aligns with the $O(1/T)$ convergence rate established in Theorem 4.1.

*Personalization-Performance Trade-off.* We use Spearman rank correlation to evaluate how well the model captured user-specific movie preferences, serving as a surrogate for the personalization-performance trade-off discussed in Theorem 5.1. Figure 2 (bottom) reveals substantial variability in Spearman correlations across clients, ranging from 0.0613 (Client 5) to 0.6126 (Client 2) in the last round, with high correlations indicating effective personalization and low correlations suggesting challenges in capturing nuanced preferences. The upward trend in median Spearman correlations (Figure 3 bottom) demonstrates the framework's increasing ability to develop personalized models aligned with individual preferences, while steadily improving global performance (Figure 3 top).

*Scaling to $K = 50$ Clients.* Similar trends were observed when scaling to $K = 50$ clients, with details provided in Appendix D.1.4.

## 6.2 Sentiment-Controlled Review Generation

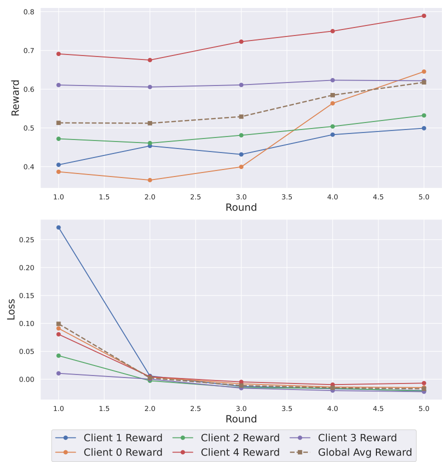
*6.2.1 Task Description and Setup.* In this task we simulate multiple movie review platforms collaborating to fine-tune a language model for sentiment-controlled text generation without sharing data. Each client represents a distinct platform with its own collection of movie reviews, introducing natural data heterogeneity. Using the IMDb dataset [22], we partitioned 50,000 reviews among $K = 5$ clients, each receiving approximately 10,000 unique reviews.

*6.2.2 Implementation (details in Appendix D.2.2).* We employed a GPT-2 model [30] fine-tuned using PPO [34] within the TRL library [39]. Clients conducted local RLHF training for 5 epochs per federated round, using Adam optimizer (learning rate $1 \times 10^{-5}$). Global aggregation used FedAvg [24] over 5 communication rounds.

*6.2.3 Human Feedback Simulation.* We simulated human feedback using a sentiment analysis model (DistilBERT [32] fine-tuned on IMDb) implemented locally on each client. The reward function combined sentiment score and language model log probability: $R_k = \lambda_k \cdot R_{\text{sentiment}} + (1 - \lambda_k) \cdot R_k^0$, where $R_{\text{sentiment}}$ is the sentiment alignment reward, $R_k^0$ is the intrinsic fluency reward and $\lambda_k \in [0, 1]$ is a client-specific parameter controlling personalization. This formulation closely aligns with the reward shaping in Equation 2 and validates Assumption 6, allowing clients to personalize the importance of sentiment alignment.

*6.2.4 Results and Analysis.* We report the results using a single random seed (42) to maintain consistency across experiments.

*Comparison with Centralized RLHF.* Figure 4b compares the average rewards and losses between centralized RLHF and FedRLHF ($K = 5$) over the total number of samples. The rewards comparison reveals that while the centralized model initially achieves slightly higher rewards, FedRLHF quickly catches up and even surpasses the centralized model's performance in later stages. This is evident from the FedRLHF reward curve (in red) consistently lying above the centralized RLHF curve (in blue) after approximately 1500 samples. This improvement arises from FedRLHF's ability to leverage diverse client data and regular model aggregation, which enhance




**(a) Global and clients performance of FedRLHF in the IMDb task.**


**(b) Sample efficiency of FedRLHF ($K = 5$) versus Centralized RLHF.**

**Figure 4: Performance evaluation of FedRLHF in comparison to centralized RLHF. (a) Tracks the rewards and losses of clients and the global performance over federation rounds. (b) Compares the sample efficiency of FedRLHF (K = 5) with centralized RLHF in terms of average rewards and losses over training samples.**

generalization and reduce overfitting compared to the centralized approach. The loss comparison shows that both approaches achieve similar loss reduction. This result corroborates the sample complexity analysis in Theorem 4.2, indicating that FedRLHF can match or even exceed centralized performance while preserving privacy and distributing computation.

*Personalization-Performance Trade-off.* The global average reward, represented by the dashed line in Figure 4a, shows steady improvement from approximately 0.52 to 0.68 over five rounds, indicating overall system convergence. To analyze the trade-off between personalization and global performance in FedRLHF, we conducted a detailed evaluation of how clients' personalized objectives affected their individual rewards over the training rounds. For each client, we randomly sampled 30 queries from their evaluation dataset at the beginning of training and kept these queries fixed throughout all rounds. Each client was assigned a different

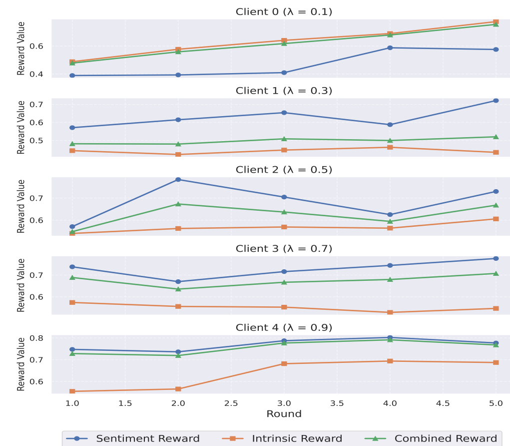



**Figure 5: Trends of intrinsic rewards, sentiment rewards, and combined rewards over communication rounds for each client. Each subplot corresponds to one client, illustrating personalization effects due to varying $\lambda_k$ values.**

personalization weight $\lambda_k$, ranging from 0.1 to 0.9. In each communication round, we supplied these 30 queries to the client's GPT model, recorded the generated responses, and calculated the corresponding average of intrinsic rewards ($R_{\text{intrinsic}}$), sentiment rewards ($R_{\text{sentiment}}$), and combined rewards ($R_k$), as shown in Figure 5.

The results reveal distinct patterns aligned with the personalization weights $\lambda_k$. Client 0 ($\lambda_0 = 0.1$) prioritizes intrinsic rewards, while Client 1 ($\lambda_1 = 0.3$) shows more balanced improvement. Client 2 ($\lambda_2 = 0.5$) exhibits clear equilibrium between sentiment and intrinsic rewards. For Client 3 ($\lambda_3 = 0.7$), sentiment rewards dominate with a steady increase, and Client 4 ($\lambda_4 = 0.9$) demonstrates the highest emphasis on sentiment rewards. As $\lambda$ increases across clients, we observe a clear shift from intrinsic to sentiment reward prioritization, with combined rewards increasingly aligning with sentiment rewards for higher $\lambda$ values.

## 7 CONCLUSION AND FUTURE WORK

In this work, we have introduced FedRLHF, a novel framework that integrates federated reinforcement learning principles with RLHF to address privacy and personalization challenges of data centralization. Our theoretical analysis provides convergence guarantees and sample complexity bounds, demonstrating stable, linear convergence. The personalization-performance trade-off analysis shows how FedRLHF balances global performance with individual client needs. Empirical evaluations on MovieLens and IMDb validate our approach, achieving results comparable to centralized RLHF while preserving privacy and enhancing personalization.

Future work will focus on enhancing FedRLHF's robustness through advanced aggregation techniques and strengthening privacy preservation by integration of formal privacy guarantees, such as differential privacy. Additionally, we aim to investigate the trade-off between communication efficiency and personalization, optimizing FedRLHF's performance by balancing communication overhead with personalized model adaptations in federated environments.

# REFERENCES


[1] Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Titouan Parcollet, and Nicholas D Lane. 2020. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390* (2020).

[2] Jalaj Bhandari and Daniel Russo. 2024. Global optimality guarantees for policy gradient methods. *Operations Research* (2024).

[3] Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, Vol. 30.

[4] Nathan Corecco, Giorgio Piatti, Luca A Lanzendörfer, Flint Xiaofeng Fan, and Roger Wattenhofer. 2024. An LLM-based Recommender System Environment. *arXiv preprint arXiv:2406.01631* (2024).

[5] Zhongxiang Dai, Flint Xiaofeng Fan, Cheston Tan, Trong Nghia Hoang, Bryan Kian Hsiang Low, and Patrick Jaillet. 2024. Federated sequential decision making: Bayesian optimization, reinforcement learning, and beyond. In *Federated Learning*. Elsevier, 257–279.

[6] Zhongxiang Dai, Yao Shu, Arun Verma, Flint Xiaofeng Fan, Bryan Kian Hsiang Low, and Patrick Jaillet. 2023. Federated Neural Bandits. In *Proceedings of the Eleventh International Conference on Learning Representations*.

[7] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*. Springer, 265–284.

[8] European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj

[9] Flint Xiaofeng Fan, Yining Ma, Zhongxiang Dai, Wei Jing, Cheston Tan, and Bryan Kian Hsiang Low. 2021. Fault-Tolerant Federated Reinforcement Learning with Theoretical Guarantee. In *Advances in Neural Information Processing Systems*. 1007–1021.

[10] Flint Xiaofeng Fan, Yining Ma, Zhongxiang Dai, Cheston Tan, Bryan Kian Hsiang Low, and Roger Wattenhofer. 2023. *Fedhql: Federated heterogeneous q-learning*. arXiv:2301.11135.

[11] Flint Xiaofeng Fan, Cheston Tan, Yew-Soon Ong, Roger Wattenhofer, and Wei-Tsang Ooi. 2024. FedRLHF: A Convergence-Guaranteed Federated Framework for Privacy-Preserving and Personalized RLHF. *arXiv preprint arXiv:2412.15538* (2024). https://arxiv.org/abs/2412.15538

[12] Koki Fujita, Shugo Fujimura, Yuwei Sun, Hiroshi Esaki, and Hideya Ochiai. 2022. Federated Reinforcement Learning for the Building Facilities. In *2022 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*. 1–6.

[13] F Maxwell Harper and Joseph A Konstan. 2015. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 4 (2015), 1–19.

[14] Wenzheng Jiang, Ji Wang, Xiongtao Zhang, Weidong Bao, Cheston Tan, and Flint Xiaofeng Fan. 2025. FedHPD: Heterogeneous Federated Reinforcement Learning via Policy Distillation. *arXiv preprint arXiv:2502.00870* (2025).

[15] Philip Jordan, Florian Grötschla, Flint Xiaofeng Fan, and Roger Wattenhofer. 2024. Decentralized Federated Policy Gradient with Byzantine Fault-Tolerance and Provably Fast Convergence. In *Proceedings of the 2024 International Conference on Autonomous Agents and Multiagent Systems*.

[16] Hamed Karimi, Julie Nutini, and Mark Schmidt. 2016. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*. Springer, 795–811.

[17] Sajad Khodadadian, Pranay Sharma, Gauri Joshi, and Siva Theja Maguluri. 2022. Federated Reinforcement Learning: Linear Speedup Under Markovian Sampling. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*.

[18] Diederik P Kingma and Jimmy Ba. 2014. *Adam: A method for stochastic optimization*. arXiv:1412.6980.

[19] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine* 37, 3 (2020), 50–60.

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

[21] Boyi Liu, Lujia Wang, and Ming Liu. 2019. Lifelong federated reinforcement learning: a learning architecture for navigation in cloud robotic systems. *IEEE Robotics and Automation Letters* 4, 4 (2019), 4555–4562.

[22] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 142–150.

[23] Hei Yi Mak, Flint Xiaofeng Fan, Luca A Lanzendörfer, Cheston Tan, Wei Tsang Ooi, and Roger Wattenhofer. 2024. CAESAR: Enhancing Federated RL in Heterogeneous MDPs through Convergence-Aware Sampling with Screening. In *Proceedings of the Sixteenth Workshop on Adaptive and Learning Agents*.

[24] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. PMLR, 1273–1282.

[25] OpenAI. 2023. ChatGPT. https://openai.com/blog/chatgpt.

[26] 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. In *Advances in Neural Information Processing Systems*.

[27] Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. 2018. Stochastic variance-reduced policy gradient. In *International conference on machine learning*. 4026–4035.

[28] 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* (2024).

[29] Jing Qiao, Zuyuan Zhang, Sheng Yue, Yuan Yuan, Zhipeng Cai, Xiao Zhang, Ju Ren, and Dongxiao Yu. 2024. BR-DeFedRL: Byzantine-Robust Decentralized Federated Reinforcement Learning with Fast Convergence and Communication Efficiency. In *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications*. 141–150. https://doi.org/10.1109/INFOCOM52122.2024.10621347

[30] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[31] Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. 2024. Group Robust Preference Optimization in Reward-free RLHF. *arXiv preprint arXiv:2405.20304* (2024).

[32] V Sanh. 2019. DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv preprint arXiv:1910.01108* (2019).

[33] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *ICML*. 1889–1897.

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

[35] Reza Shokri and Vitaly Shmatikov. 2015. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1310–1321.

[36] Virginia Smith, Dmitriy Smola, and Ameet Talwalkar. 2017. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, Vol. 30.

[37] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize with human feedback. In *NeurIPS*, Vol. 33. 3008–3021.

[38] United States Congress. 1996. Health Insurance Portability and Accountability Act of 1996. https://www.hhs.gov/hipaa/index.html.

[39] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformer Reinforcement Learning.

[40] Han Wang, Sihong He, Zhili Zhang, Fei Miao, and James Anderson. 2024. Momentum for the Win: Collaborative Federated Reinforcement Learning across Heterogeneous Environments. *arXiv preprint arXiv:2405.19499* (2024).

[41] Xiaofei Wang, Chenyang Wang, Xiuhua Li, Victor CM Leung, and Tarik Taleb. 2020. Federated deep reinforcement learning for Internet of Things with decentralized cooperative edge caching. *IEEE IoT-J* 7, 10 (2020), 9441–9455.

[42] Christopher John Cornish Hellaby Watkins. 1989. *Learning from delayed rewards*. PhD thesis, University of Cambridge England.

[43] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3-4 (1992), 229–256.

[44] Jiin Woo, Gauri Joshi, and Yuejie Chi. 2023. The Blessing of Heterogeneity in Federated Q-Learning: Linear Speedup and Beyond. In *Proceedings of the 40th International Conference on Machine Learning*.

[45] Jiin Woo, Laixi Shi, Gauri Joshi, and Yuejie Chi. 2024. Federated offline reinforcement learning: Collaborative single-policy coverage suffices. *arXiv preprint arXiv:2402.05876* (2024).

[46] Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. 2018. *Byzantine-robust distributed learning: Towards optimal statistical rates*. arXiv:1803.01498.

[47] Shuai Yu, Xu Chen, Zhi Zhou, Xiaowen Gong, and Di Wu. 2020. When deep reinforcement learning meets federated learning: Intelligent multi-timescale resource management for multi-access edge computing in 5G ultra dense network. *IEEE Internet of Things Journal* (2020).

[48] Rui Yuan, Robert M Gower, and Alessandro Lazaric. 2022. A general sample complexity analysis of vanilla policy gradient. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3332–3380.

[49] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, and Dario Amodei. 2019. Fine-Tuning Language Models from Human Preferences. *arXiv preprint arXiv:1909.08593* (2019).

[50] James Zou, Yu Lu, Balakrishnan Narayanaswamy, and Devi Parikh Agarwal. 2019. Reinforcement learning to optimize long-term user engagement in recommender systems. In *Proceedings of the 25th ACM SIGKDD KDD*. 2810–2818.