Privacy-Preserving Reinforcement Learning from Human Feedback via Decoupled Reward Modeling

Abstract

Preference-based fine-tuning has become an important component in training large language models, and the data used at this stage may contain sensitive user information. A central question is how to design a differentially private pipeline that is well suited to the distinct structure of reinforcement learning from human feedback. We propose a privacy-preserving framework that imposes differential privacy only on reward learning and derives the final policy from the resulting private reward model. Theoretically, we study the suboptimality gap and show that privacy contributes an additional additive term beyond the usual non-private statistical error. We also establish a minimax lower bound and show that the dominant term changes with sample size and privacy level, which in turn characterizes regimes in which the upper bound is rate-optimal up to logarithmic factors. Empirically, synthetic experiments confirm the scaling predicted by the theory, and experiments on the Anthropic HH-RLHF dataset using the Gemma-2B-IT model show stronger private alignment performance than existing differentially private baseline methods across privacy budgets.

Keywords: Large Language Models, Preference Fine-tuning, Differential Privacy, Reward Modeling, Sample Complexity.

Mathematics Subject Classification (2020): 62XXX

1 Introduction

Large language models (LLMs) are increasingly used as general-purpose tools across a growing range of downstream and domain-specific applications, including medicine, finance, law, and science (Nie et al., 2024; Qin and Sun, 2024; Zheng et al., 2025; Zhou et al., 2026). In practice, adapting a pretrained model to such settings typically relies on post-training or fine-tuning, which has become a standard part of modern LLM deployment (Ouyang et al., 2022; Bai et al., 2022; Chia et al., 2025). Preference-based LLM fine-tuning is one widely used form of this adaptation. Such fine-tuning is commonly carried out through reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022), where comparative feedback is used to improve a policy when reward feedback is not directly observed.

In particular, a typical LLM alignment pipeline begins with a pretrained model, obtains a reference policy through supervised fine-tuning (SFT) on expert-written or carefully curated instruction-response pairs, and then refines that policy using pairwise preferences over candidate responses, typically through RLHF. Because SFT data are costly to scale and may not fully capture the nuanced judgments needed for alignment, preference-based fine-tuning has become a common extension beyond SFT (Chia et al., 2025). Figure 1 illustrates this canonical

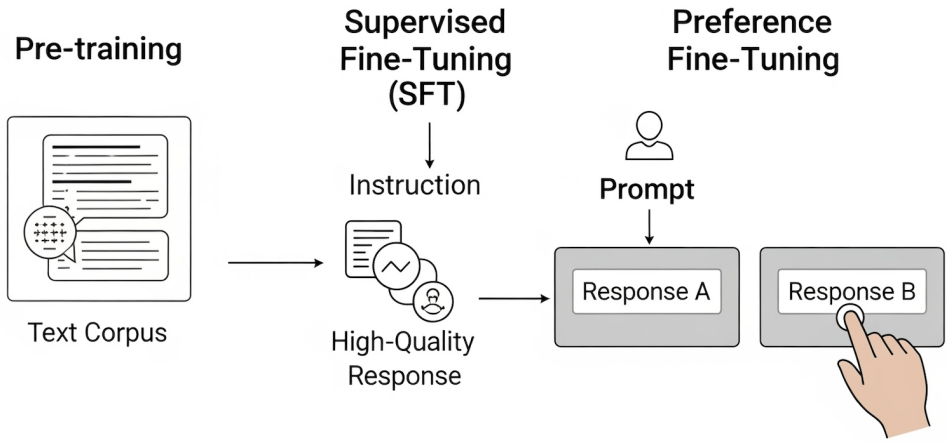


Figure 1: A typical large language model adaptation pipeline. We focus on privacy during the preference fine-tuning stage, where sensitive user interactions can be directly reflected in training records.

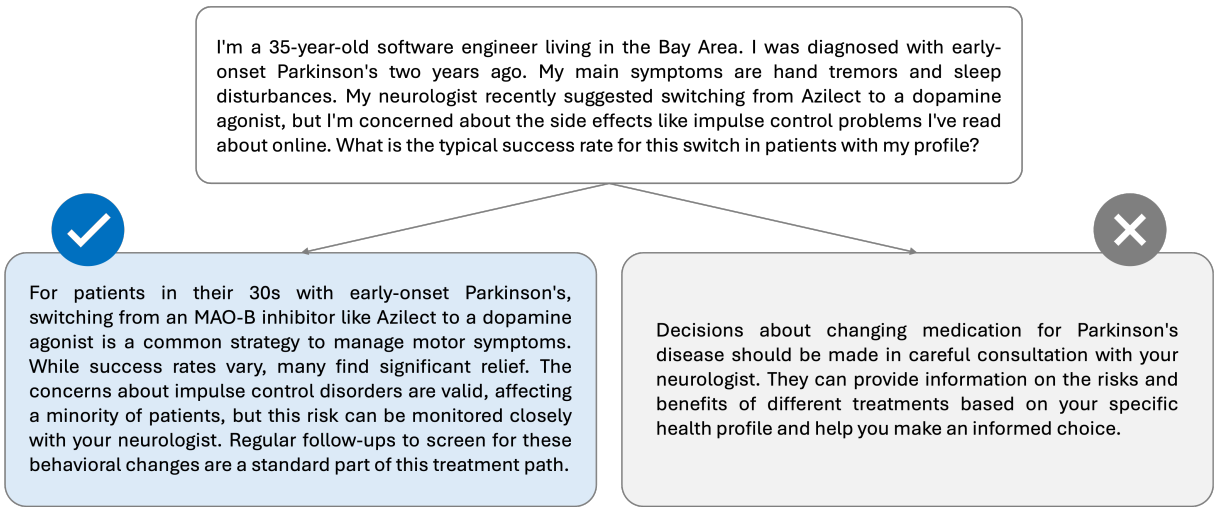


Figure 2: A sensitive interaction record. Even without direct identifiers, prompts can contain quasi-identifiers whose combination may re-identify an individual, motivating tuple-level protection of (x_i, a_i^1, a_i^2, y_i).

RLHF pipeline. Modern preference-based alignment also includes direct formulations such as direct preference optimization (DPO), which bypass explicit reward-model training and instead optimize the policy directly from preference data (Rafailov et al., 2024).

On the other hand, this preference-based fine-tuning pipeline can raise privacy concerns. Modern models are susceptible to training-data extraction attacks (Carlini et al., 2021; Nasr et al., 2023) and membership inference attacks (Shokri et al., 2017; Wu and Cao, 2025), which can reveal whether a user's data was used for training and, in some cases, expose training examples. In preference-based LLM fine-tuning beyond SFT, the training signal is not merely a scalar label, but a full interaction tuple (x_i, a_i^1, a_i^2, y_i), where x_i is the user prompt, (a_i^1, a_i^2) are candidate responses, and y_i is the associated preference label. Even when explicit identifiers are absent, the prompt x_i may contain sensitive or potentially identifying context, and the derived responses and label may also reveal private user information. Figure 2 illustrates such an example. These concerns call for principled protection against leakage at the tuple level.

Differential privacy (DP) (Dwork et al., 2006) provides a widely used formal framework for privacy-preserving data analysis and machine learning. A central question is what object should be protected by the privacy definition. Much of the existing literature adopts label-DP (Chowdhury et al., 2024; Zhang et al., 2025; Teku et al., 2025; Wu et al., 2025), which protects only the preference label and is therefore most naturally aligned with protecting the privacy of the preference-label annotator. Our focus is different. In the settings we study, the relevant target is the end-user, so privacy should apply to the full interaction tuple rather than to y_i alone. This motivates tuple-level privacy for (x_i, a_i^1, a_i^2, y_i).

A key distinction from standard RL is that RLHF does not receive rewards directly from the environment. Instead, it must infer a latent reward signal from preference data, which introduces an additional reward-learning layer into the pipeline. Under DP, this distinction is consequential. Privacy can now enter at reward estimation, at downstream policy optimization, or at both stages. If privacy is imposed naively on such a multi-stage pipeline, noise can accumulate across stages, or the data or privacy budget must be divided across them. In this sense, the reward-learning layer in RLHF is not only a modeling feature but also a central channel through which privacy constraints affect utility.

This raises a second central question beyond what should be protected. One must also decide where privacy should enter the alignment pipeline so that downstream utility is preserved as much as possible. Existing directions already illustrate several possibilities, including introducing DP to DPO (Chen et al., 2025) and introducing DP to each stage of RLHF (Wu et al., 2024). In this sense, existing methods largely arise by applying DP to existing non-private alignment pipelines. The remaining question is whether RLHF admits a privacy-aware design that uses its own structure more directly.

In this work, we take a different route. Rather than adding DP to each step of an existing alignment pipeline, we use the two-stage structure of RLHF itself to guide where privacy should enter. We develop a tuple-level private framework that places privacy on reward learning and treats downstream policy improvement as post-processing. We then study the resulting privacy–utility tradeoff through theory and numerical experiments.

1.1 Our Contributions

Our contributions are summarized as follows.

- **Methodology.** We propose a private RLHF framework that exploits a structural feature specific to RLHF, rather than simply adding DP to an existing alignment pipeline. Unlike standard RL, where the environment directly provides reward, RLHF first uses preference data to train a reward model and then updates a reference policy using that learned reward. Our framework places DP only on the reward-learning stage and treats downstream policy as post-processing of the resulting private reward model. Placing privacy on reward learning creates a buffer between DP noise and the downstream policy. In LLM fine-tuning, this avoids privatizing the policy itself, which can yield responses worse than the reference policy, and instead uses the private reward model only to re-rank reference-generated responses, without additional privacy cost.

- **Theory.** We study the suboptimality gap of the policy induced by our framework and show that privacy enters as an additional additive term on top of the usual non-private statistical error. We also establish, to the best of our knowledge, the first minimax lower bound for this private RLHF problem. A key difficulty is that the dominant lower-bound term changes with the balance between sample size and privacy budget. We therefore characterize the regimes in which our method is rate-optimal up to logarithmic factors, including the case where the sample size is sufficiently large relative to the privacy scale for the optimal privacy-dominated rate to emerge.

- **Empirical validation.** We validate the framework on synthetic preference-learning experiments and on an LLM alignment task based on the Anthropic HH-RLHF dataset (Bai et al., 2022) using the Gemma-2B-IT model (Team et al., 2024). The synthetic results confirm the theoretically predicted scaling with sample size, privacy budget, and dimension, and show that our method substantially reduces the incidence of policies that underperform the reference policy relative to existing DP baseline methods. In the LLM experiment, our method demonstrates stronger private alignment performance than existing DP baseline methods across all privacy budgets considered.

1.2 Related Work

We briefly review prior work on preference-based policy learning and its theoretical foundations, as well as DP for alignment.

Preference-based policy learning and KL-regularized preference optimization. One of the most widely used applications of RLHF is preference fine-tuning for LLMs (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022; Ye et al., 2025). More recently, direct alignment methods have gained prominence by optimizing a closed-form objective that avoids explicit reward-model training and RL in the loop (Rafailov et al., 2024; Garg et al., 2025). From a theoretical viewpoint, a growing line of work develops finite-sample guarantees of KL-regularized preference optimization and iterative procedures, clarifying when logged preference data suffices for reliable policy improvement (Xiong et al., 2024; Ye et al., 2024; Xiong et al., 2024; Song et al., 2024; Zhao et al., 2024). Several recent works also revisit the role of the KL regularizer itself—including its interpretation, limitations, and variants (Huang et al., 2024; Aminian et al., 2025; Liu et al., 2025a; Li et al., 2026). Our work complements this theory by quantifying how DP affects the KL-regularized preference optimization.

DP and privacy-preserving alignment. DP is a *de facto* standard in privacy-preserving framework in machine learning (Dwork et al., 2006). A large literature studies optimal accuracy guarantees under DP for empirical risk minimization and stochastic convex optimization (Dwork et al., 2014; Chaudhuri et al., 2011; Kifer et al., 2012; Bassily et al., 2014; Wang et al., 2017; Bassily et al., 2019). Among many, the most widely used optimization method is noisy gradient-based training, popularized by DP-SGD and closely related mechanisms (Dwork et al., 2006; Abadi et al., 2016; Bu et al., 2020). This line of work has also motivated refined privacy accounting frameworks and composition results that tighten end-to-end privacy loss in iterative training (Kairouz et al., 2015; Mironov, 2017; Bun and Steinke, 2016; Dong et al., 2022).

Building on these foundations, recent work has begun to study DP for preference-based

alignment and RLHF. A first key distinction concerns the unit of protection. Much of the recent literature studies label-DP, which protects only the preference label while treating the prompt and candidate responses as public or non-sensitive. This formulation is naturally aligned with protecting annotator feedback, but it does not address leakage from the text content itself. This perspective underlies several recent works on private DPO and related offline alignment objectives (Zhang et al., 2025; Zhou et al., 2025a,b; Teku et al., 2025). In contrast, our focus is tuple-level privacy, where the prompt, candidate responses, and preference outcome may all carry sensitive information.

A second distinction concerns how privacy is incorporated into the alignment pipeline. Most existing approaches begin with a non-private alignment framework and then introduce a DP mechanism on top of it. When the target is label-DP, a standard way to privatize the labels is randomized response (Warner, 1965), and several recent works follow this route (Zhang et al., 2025; Zhou et al., 2025a,b; Teku et al., 2025). When the target is tuple-level DP, privacy is typically enforced through noisy gradient-based training such as DP-SGD, including private adaptations of DPO and related RLHF procedures (Chen et al., 2025; Wu et al., 2024, 2025; Chowdhury et al., 2024; Korkmaz and Brown-Cohen, 2024). Our contribution differs from these directions both methodologically and theoretically. We study a framework tailored to the distinct structure of RLHF, with privacy imposed only on reward learning, and we analyze the resulting suboptimality gap together with minimax lower bounds that characterize when the corresponding rates are optimal up to logarithmic factors.

1.3 Paper organization and notation

The remainder of the paper is organized as follows. Section 2 introduces the preliminaries for our work, including DP and RLHF. Section 3 presents the proposed decoupled framework. In section 4, we develop the theoretical analysis, deriving both upper and lower bounds on the suboptimality gap of the induced policy and identifying regimes in which these bounds match up to logarithmic factors. Section 5 reports numerical studies on both synthetic examples and LLM fine-tuning experiments. Section 6 concludes with a discussion of the main implications, limitations, and directions for future work. Most proofs are deferred to the Supplementary Material Section D, while straightforward arguments are given in the main text.

For notation, we use standard asymptotic order notation. For two positive sequences a_n and b_n, $a_n = O(b_n)$ means that a_n/b_n is uniformly bounded, and $a_n = \Omega(b_n)$ means that a_n/b_n is bounded away from zero. We write $a_n = \widetilde{O}(b_n)$ and $a_n = \widetilde{\Omega}(b_n)$ when the corresponding relations hold up to polylogarithmic factors.

2 Preliminaries

In this section, we establish the background for our work. We begin by formally outlining the standard pipelines for preference fine-tuning, then introduce the background on DP.

2.1 Reinforcement Learning from Human Feedback

A widely used template for learning from preferences is KL-regularized reinforcement learning from human feedback (RLHF). The key object is a reference policy π_0, and the goal is to improve decision making while preventing the updated policy from drifting too far from π_0 or overfitting to a limited preference dataset. This viewpoint is especially prominent in large language model alignment, where preference fine-tuning is a standard stage after a strong base policy is obtained.

We consider an offline preference dataset originally recorded as $\mathcal{D} = \{(x_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$, where $x_i \in \mathcal{X}$ denotes a context, such as a prompt in LLM applications, $a_i^1, a_i^2 \in \mathcal{A}$ are two candidate actions, and y_i indicates which candidate is preferred. For the theoretical development, it is more convenient to rewrite each record in ordered form as (x_i, a_i^w, a_i^l), where $a_i^w \in \mathcal{A}$ and $a_i^l \in \mathcal{A}$ denote the preferred and non-preferred actions, respectively. A conventional modeling assumption is then the Bradley–Terry preference model (Bradley and Terry, 1952).

Definition 1 (Bradley–Terry Model)**.** *For a context x and a preferred/non-preferred pair (a^w, a^l), the probability that a^w is preferred to a^l is modeled as*

$$\mathbb{P}(a^w \succ a^l \mid x) = \frac{\exp(r^*(x, a^w))}{\exp(r^*(x, a^w)) + \exp(r^*(x, a^l))} = \sigma\big(r^*(x, a^w) - r^*(x, a^l)\big),$$

where $\sigma(t) = (1 + e^{-t})^{-1}$ is the sigmoid function.

In KL-regularized RLHF, the reference policy π_0 is treated as a strong baseline that encodes prior knowledge and safe behavior, and KL regularization controls the magnitude of the preference-driven update. In the LLM preference fine-tuning pipeline, a common choice of π_0 is a model that has already been trained on a large supervised instruction-following dataset, often referred to as supervised fine-tuning, or SFT (Ouyang et al., 2022). This choice reflects the practical role of π_0 as an anchor to a broadly competent response distribution, while preference data provide an additional signal that refines behavior without requiring the policy to relearn basic capabilities from scratch.

Under Definition 1, the standard reward-modeling phase estimates a parametric reward function $r_\theta \in \mathcal{R} = \{r_\theta : \theta \in \Theta\}$ by maximizing the log-likelihood

$$\hat{\theta} = \operatorname*{argmax}_{\theta \in \Theta} \sum_{i=1}^{n} \log \sigma\left(r_\theta(x_i, a_i^w) - r_\theta(x_i, a_i^l)\right). \tag{1}$$

Once a reward estimator $\hat{r}$ is obtained, the goal is to derive a policy that achieves high reward while remaining close to a reference policy π_0, which is often the SFT model in LLM pipelines. For any reward function r on $\mathcal{X} \times \mathcal{A}$, we evaluate a policy π by the KL-regularized value

$$V_\eta(\pi; r) = \mathbb{E}_{x \sim d_0}\left[\mathbb{E}_{a \sim \pi(\cdot|x)}\big[r(x, a)\big] - \frac{1}{\eta}\mathrm{KL}\big(\pi(\cdot \mid x) \,\|\, \pi_0(\cdot \mid x)\big)\right], \tag{2}$$

where d_0 denotes the context distribution, and $\eta > 0$ controls the strength of regularization. Let $\pi_r^\eta \in \arg\max_\pi V_\eta(\pi; r)$ denote the KL-regularized optimizer under r. The closed form solution is as follows:

Lemma 2 (Policy Improvement Oracle)**.** *For a fixed context x, reward r, and reference policy π_0, any maximizer $\pi_r^\eta \in \arg\max_\pi V_\eta(\pi; r)$ satisfies*

$$\pi_r^\eta(a \mid x) = \frac{1}{Z_r(x)} \pi_0(a \mid x) \exp(\eta\, r(x, a)), \tag{3}$$

where $Z_r(x) = \mathbb{E}_{a\sim\pi_0(\cdot|x)}\big[\exp(\eta r(x, a))\big]$.

Although π_r^η admits a closed-form expression, exact sampling from this policy is typically infeasible when the action space is large. The expression involves a normalization term that aggregates $\exp(\eta r(x, a))$ over the action space, which is computationally prohibitive in LLM settings where actions correspond to long token sequences. This motivates practical methods that approximate the KL-regularized optimizer through parameterized learning.

A common approach is to optimize a parameterized policy with reinforcement learning algorithms such as proximal policy optimization (PPO) (Schulman et al., 2017), which targets the KL-regularized objective while avoiding explicit normalization over $\mathcal{A}$. Since PPO does not appear explicitly in the main development, we defer a brief background discussion to Appendix A. An alternative is Direct Preference Optimization (DPO) (Rafailov et al., 2024), which leverages the identity implied by (3). In particular, for π_r^η one can write

$$r(x, a) = \frac{1}{\eta} \log \frac{\pi_r^\eta(a \mid x)}{\pi_0(a \mid x)} + \frac{1}{\eta} \log Z_r(x).$$

Substituting this relation into the pairwise preference likelihood (1) eliminates the unknown $Z_r(x)$ and yields a supervised objective over policy parameters

$$\mathcal{L}_{\text{DPO}}(\theta; \mathcal{D}, \pi_0) = -\sum_{(x,a^w,a^l)\in\mathcal{D}} \log \sigma\left(\frac{1}{\eta} \log \frac{\pi_\theta(a^w \mid x)}{\pi_0(a^w \mid x)} - \frac{1}{\eta} \log \frac{\pi_\theta(a^l \mid x)}{\pi_0(a^l \mid x)}\right). \tag{4}$$

The primary appeal of DPO and its variants (Azar et al., 2024; Ethayarajh et al., 2024; Xu et al., 2024; Meng et al., 2024) lies in their ability to avoid the computational complexities and training instabilities of RL frameworks. By formulating the process as a single-stage supervised learning, these methods eliminate the need for iterative sampling and complex hyperparameter tuning required by RL-based optimization. However, as we discuss in Section 3, directly imposing DP on the policy parameters in (4) introduces fundamental challenges.

Our analysis focuses on the policy quality under the true reward r^*. Let $\pi_{r^*}^\eta \in \arg\max_\pi V_\eta(\pi; r^*)$. We quantify the η-regularized suboptimality of a candidate policy π by

$$\Delta_\eta(\pi) = V_\eta(\pi_{r^*}^\eta; r^*) - V_\eta(\pi; r^*). \tag{5}$$

In the proposed decoupled framework, π will be induced by a private reward estimate, and our theory controls $\Delta_\eta(\pi)$ in finite samples.

2.2 Differential Privacy

Informally, a mechanism M is DP if the distribution of its output $M(D)$ is nearly indistinguishable from that of $M(D')$ for any adjacent dataset D', where D and D' differ by a single entry.

We formalize the notion of adjacent as their Hamming distance is one.

Among various characterizations of DP–such as Rényi DP (Mironov, 2017) or Gaussian DP (Dong et al., 2022)–aimed at achieving tighter privacy accounting, the following (ε, δ)-DP remains the most prevalent and standard definition, and thus we also adopt the following:

Definition 3 ((ε, δ)-Differential Privacy (Dwork et al., 2006))**.** *A mechanism M is (ε, δ)-differentially private if, for any two adjacent D, D' and for any measurable event E,*

$$\mathbb{P}(M(D) \in E) \leq e^{\varepsilon}\, \mathbb{P}(M(D') \in E) + \delta.$$

When $\delta = 0$, the mechanism M is said to satisfy ε-differential privacy.

DP protects individuals by requiring that an algorithm produce similar output distributions on adjacent datasets that differ in a single record. This similarity limits the influence of any one data point on what is released, so an observer cannot reliably infer whether that record was included. The (ε, δ) definition measures similarity through a log-likelihood ratio control between the two output distributions for every measurable event. The parameter ε bounds the magnitude of this log-likelihood ratio, while δ allows a small probability mass on which the bound may be violated, which can be viewed as an ε-DP guarantee holding with probability at least $1 - \delta$. Smaller values of ε and δ therefore correspond to stronger privacy protection, and δ is typically chosen to be extremely small in practice, since it quantifies the probability of a rare failure event the privacy protection may not hold.

A key appeal of DP is that its guarantees are accompanied by explicit parameters that can be tracked across an entire pipeline, which makes it possible to quantify the overall level of protection. In particular, DP enjoys the following properties:

- **Post-processing:** If M is (ε, δ)-DP, then for any mapping Proc independent of the data, the post-processed mechanism $\mathrm{Proc} \circ M$ is also (ε, δ)-DP.
- **Sequential composition (basic):** Let $M_1 : \mathcal{X} \to \mathcal{Y}_1$ and $M_i : \mathcal{X} \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{i-1} \to \mathcal{Y}_i$, for $i = 2, \ldots, k$. Let $M(D) := (M_1(D),\, M_2(D, y_1),\, \ldots,\, M_k(D, y_1, \ldots, y_{k-1}))$ denote the joint mechanism, where the outputs are generated recursively. If, for every fixed previous output history, each step is $(\varepsilon_i, \delta_i)$-DP, then M is $\left(\sum_{i=1}^{k} \varepsilon_i,\ \sum_{i=1}^{k} \delta_i\right)$-DP.
- **Parallel composition:** Let D be partitioned into disjoint subsets $D_1, \ldots, D_k$, and let M_i be an $(\varepsilon_i, \delta_i)$-DP mechanism applied only to D_i. Define the joint mechanism $M(D) := (M_1(D_1), \ldots, M_k(D_k))$. Then M is $\left(\max_{1 \leq i \leq k} \varepsilon_i,\ \max_{1 \leq i \leq k} \delta_i\right)$-DP.

We present the basic sequential composition rule only for illustration, and the main point is that privacy parameter accumulates across multiple data-dependent releases. In practice, one can obtain tighter accounting by using refined composition theorems and privacy definitions designed for sharper tracking of cumulative loss, such as Rényi DP or Gaussian DP. We do not pursue these refinements here since our focus is on the pipeline-level design and on policy-quality guarantees, rather than on the tightest possible accounting constants.

Constructing a DP mechanism typically involves injecting noise, calibrated to the *sensitivity*, the maximum impact of a single data point on the output. In machine learning where gradient-based optimization is standard, the contribution of an individual data is its respective gradient.

Therefore, a natural pathway to DP is to inject noise into the gradient updates, leading to the DP-stochastic gradient descent (DP-SGD) (Abadi et al., 2016).

Concretely, let $\ell(\theta; z)$ denote a per-example loss function and $\Theta \subset \mathbb{R}^d$ be a parameter space. A standard minibatch DP-SGD update is defined as:

$$\theta_{t+1} = \Pi_\Theta \left(\theta_t - \eta_t \left(\frac{1}{|B_t|} \sum_{i \in B_t} \text{clip}(\nabla_\theta \ell(\theta_t; Z_i), C) + \xi_t \right) \right), \quad \xi_t \sim \mathcal{N}(0, \sigma_{\text{DP}}^2 C^2 I_d), \tag{6}$$

where $B_t \subset [n]$ is a minibatch, Π_Θ denotes the projection onto the set Θ, and $\text{clip}(g, C) := g \cdot \min\{1, C/\|g\|_2\}$ scales each gradient to ensure a uniform ℓ_2-sensitivity bound C. While sensitivity may not be inherently bounded or analytically tractable–particularly for black-box models– this gradient clipping ensures that the sensitivity is always bounded by C, which is then used to calibrate the noise. As such, gradient clipping serves as a key tool that provides the flexibility to satisfy DP for arbitrary differentiable architectures.

Other optimization approaches are primarily distinguished by the stage at which noise is introduced, such as objective perturbation or output perturbation (Chaudhuri et al., 2011). For large-scale tasks, however, DP-SGD remains the *de facto* standard and the most widely used scalable paradigm in practice, supported by implementations such as the `Opacus` library in Python (Yousefpour et al., 2021).

3 Proposed Method

3.1 Motivation: Challenges in Private Policy Optimization

The prevailing paradigms in finetuning, such as DPO and PPO, optimize the policy parameters π_θ directly. We defer a more detailed description of PPO to Appendix B.1, and focus here on the common issue that arises when DP is imposed on policy optimization.

The fundamental difficulty stems from the inherent incompatibility between the unbounded nature of policy gradients and the sensitivity required by DP. In frameworks like PPO or DPO, the gradient involves the score function

$$\nabla_\theta \log \pi_\theta(a|x) = \frac{\nabla \pi_\theta(a|x)}{\pi_\theta(a|x)},$$

which lacks a uniform upper bound since $\pi_\theta(a|x)$ can be arbitrarily close to zero, making the ratio arbitrarily large. As these policy gradients can be arbitrarily large, the aforementioned clipping during the gradient updates are necessary.

Example 4 (DPO)**.** *Consider the DPO objective in* (4)*. For a single preference pair* (x, a^w, a^l)*, define*

$$z_\theta(x, a^w, a^l) := \frac{1}{\eta} \log \frac{\pi_\theta(a^w|x)}{\pi_0(a^w|x)} - \frac{1}{\eta} \log \frac{\pi_\theta(a^l|x)}{\pi_0(a^l|x)}.$$

The per-pair loss is $\ell(\theta) = -\log \sigma(z_\theta(x, a^w, a^l))$*. Since* π_0 *is fixed, the gradient is*

$$\nabla_\theta \ell(\theta) = -(1 - \sigma(z_\theta)) \nabla_\theta z_\theta = -\frac{1}{\eta}(1 - \sigma(z_\theta)) \Big(\nabla_\theta \log \pi_\theta(a^w \mid x) - \nabla_\theta \log \pi_\theta(a^l \mid x) \Big).$$

This expression depends on the score function $\nabla_\theta \log \pi_\theta(a \mid x)$, which generally admits no uniform bound because $\pi_\theta(a \mid x)$ can be arbitrarily small. As a result, per-example gradients can be arbitrarily large under standard policy parameterizations, and DP training therefore requires explicit sensitivity control through per-example gradient clipping.

Example 5 (PPO-style policy optimization)**.** *A similar issue arises for PPO-style policy optimization. At a high level, PPO uses an advantage signal to indicate whether a sampled action should become more likely or less likely under the updated policy, while constraining the update so that the policy does not move too far in a single step. Here the advantage A is a scalar quantity summarizing how favorable action a is at context x relative to a baseline. The PPO-style ratio*

$$\rho_\theta(x, a) := \frac{\pi_\theta(a \mid x)}{\pi_{\text{ref}}(a \mid x)} = \exp\Big(\log \pi_\theta(a \mid x) - \log \pi_{\text{ref}}(a \mid x) \Big)$$

measures how much the updated policy reweights action a relative to a fixed reference policy π_{ref}. Given a parameter $\varepsilon_{\text{clip}} > 0$, the clipped surrogate takes the form

$$\ell_{\text{PPO}}(\theta) = -\min\Big\{ \rho_\theta(x, a)\, A,\ \text{clip}\big(\rho_\theta(x, a), 1 - \varepsilon_{\text{clip}}, 1 + \varepsilon_{\text{clip}}\big)\, A \Big\}.$$

Whenever the unclipped branch is active, differentiating with respect to θ yields

$$\nabla_\theta \rho_\theta(x, a) = \rho_\theta(x, a)\, \nabla_\theta \log \pi_\theta(a \mid x),$$

so the gradient again depends on the score function $\nabla_\theta \log \pi_\theta(a \mid x)$. Under standard policy parameterizations, this score function generally admits no uniform bound because $\pi_\theta(a \mid x)$ can be arbitrarily small. Thus, PPO clipping controls the objective through the policy ratio, but it does not remove the need for explicit per-example gradient clipping when DP is imposed.

Clipping plays two roles in private training. It enforces a sensitivity bound, but it can also distort optimization when many per-example gradients exceed the clipping threshold. Let $F(\theta) = \mathbb{E}_z[\ell(\theta; z)]$ and define the minibatch gradient $\widehat{\nabla} F_t(\theta)$. Under standard sampling assumptions, $\widehat{\nabla} F_t(\theta)$ is an unbiased estimator of $\nabla F(\theta)$, whereas replacing per-example gradients by $\text{clip}(\nabla_\theta \ell(\theta; z), C)$ generally introduces bias.

This distinction is especially relevant in policy optimization, where per-example gradients can be heavy-tailed due to score-function terms, so clipping may be frequently active. In contrast, when the learning target admits a uniform per-example gradient bound, one can choose C to match this bound so that clipping is rarely active and introduces negligible distortion. In that case, privacy is enforced primarily through additive noise calibrated to the sensitivity bound, which preserves mean-zero updates while inflating variance.

To address these challenges, we place the privacy mechanism on reward learning rather than on policy optimization. The reward-modeling objective is typically better conditioned, and in many LLM pipelines it is implemented by freezing a pretrained backbone and training a lightweight head on top of its representations (Evci et al., 2022; Houlsby et al., 2019; Hu et al., 2022). Let $\phi(x, a) \in \mathbb{R}^d$ denote the fixed representation of a context action pair, and consider a linear-head reward model

$$r_\theta(x, a) = \langle \phi(x, a), \theta \rangle, \tag{7}$$

with $\theta \in \Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \leq R\}$. The key point is that sensitivity control becomes transparent once the per-example loss is Lipschitz in θ. A convenient way to ensure this is to keep the representation norm bounded, for instance by applying a final normalization step on the backbone features. In modern Transformers, LayerNorm and RMSNorm already stabilize feature scales in practice (Ba et al., 2016; Zhang and Sennrich, 2019; Zheng et al., 2024), and an explicit final projection or normalization can enforce a deterministic bound $\sup_{x,a} \|\phi(x,a)\|_2 \leq L$.

Under the Bradley–Terry model, each observation $z = (x, a^w, a^\ell)$ induces a convex loss $\ell(\theta; z)$ whose gradient has the form

$$\nabla_\theta \ell(\theta; z) = \alpha(\theta; z)\big(\phi(x, a^w) - \phi(x, a^\ell)\big), \qquad \alpha(\theta; z) \in [0, 1]. \tag{8}$$

Therefore, we have

$$\|\nabla_\theta \ell(\theta; z)\|_2 \leq \|\phi(x, a^w) - \phi(x, a^\ell)\|_2 \leq \|\phi(x, a^w)\|_2 + \|\phi(x, a^\ell)\|_2 \leq 2L. \tag{9}$$

This yields a uniform per-example gradient bound, which provides a clean sensitivity control for private optimization. This suggests that per-example gradients in head-only reward learning are typically better behaved than in policy optimization. As a result, when the same per-example gradient clipping norm C is used to control DP sensitivity, clipping tends to be less frequently active in reward learning than in policy optimization. In that sense, privacy in reward learning is driven more by calibrated noise than by clipping-induced distortion.

3.2 Proposed Framework: Private Reward-Based Alignment

We propose a decoupled framework that concentrates the privacy expenditure solely on estimating the reward structure and derives the final decision rule via post-processing.

As outlined in Algorithm 1, the procedure consists of two stages: (i) learning a differentially private reward model $\tilde{r}$ from the full preference dataset $\mathcal{D}$, and (ii) producing an aligned action/response induced by $\tilde{r}$ without any further access to $\mathcal{D}$.

In the first stage, we treat reward learning as a single empirical risk minimization problem on $\mathcal{D}$. We state the framework at the level of the resulting (ε, δ)-DP guarantee, rather than specifying an explicit closed-form calibration for the added Gaussian noise. In practice, the calibration of DP-SGD depends on various factors, including the clipping norm, sampling scheme, number of epochs, and privacy accounting method (Abadi et al., 2016; Bu et al., 2020). Since our focus is on the pipeline design and on the privacy–utility trade-off at the level of the final guarantee, rather than on accountant-specific calibration formulas, we do not make the noise level explicit here. In the theoretical development in Section 4, we work with the projected noisy stochastic-gradient procedure of Bassily et al. (2014) as a concrete DP-SGD instantiation, since it delivers the strongly-convex excess-risk rate used in our analysis. In the experiments, the corresponding calibration is handled by `Opacus`, which takes the target privacy parameters together with the training configuration and internally performs privacy accounting to determine the required noise level; implementation details are deferred to Appendix B.4 and Table 3.

In the second stage, the KL-regularized target policy induced by $\tilde{r}$ is defined by the Gibbs

Algorithm 1 Differentially Private Reward-Based Alignment

Input: Preference dataset $\mathcal{D} = \{(x_i, a_i^w, a_i^l)\}_{i=1}^n$, privacy budget (ε, δ), KL-regularization parameter η, reference policy π_0, sampling budget N (optional).
Step 1: Private Reward Learning
Learn a private reward model $\tilde{r}(x, a)$ by an (ε, δ)-DP mechanism (e.g., DP-SGD).
Step 2: Policy Derivation (Inference)
if normalization constant $Z_{\tilde{r}}(x) = \mathbb{E}_{a\sim\pi_0}[\exp(\eta\tilde{r}(x, a))]$ is computable **then**
  Exact Inference
  Construct the optimal policy in closed form:

$$\pi_{\tilde{r}}^{\eta}(a|x) = \frac{1}{Z_{\tilde{r}}(x)}\pi_0(a|x)\exp\left(\eta\tilde{r}(x, a)\right).$$

else
  Approximate Inference via Best-of-N
  For a given context x, sample N candidates from reference: $\{a^{(1)}, \ldots, a^{(N)}\} \sim \pi_0(\cdot|x)$.
  Select the candidate maximizing the private reward:

$$a^* = \operatorname*{argmax}_{j\in\{1,\ldots,N\}} \tilde{r}(x, a^{(j)}).$$

  Define the policy output as the Dirac mass on a^*.
end if
Output: Privately aligned policy $\pi_{\tilde{r}}^{\eta}$ or action a^*.

form

$$\pi_{\tilde{r}}^{\eta}(a \mid x) = \frac{\pi_0(a \mid x)\exp\{\eta\,\tilde{r}(x, a)\}}{Z_{\tilde{r}}^{\eta}(x)},$$

where $Z_{\tilde{r}}^{\eta}(x) := \sum_{a'\in\mathcal{A}} \pi_0(a' \mid x)\exp\{\eta\,\tilde{r}(x, a')\}$. The implementation depends on the tractability of the partition function $Z_{\tilde{r}}^{\eta}(x)$. When $Z_{\tilde{r}}^{\eta}(x)$ is tractable (e.g., finite action sets or structured settings where normalization is feasible), one can explicitly construct and sample from $\pi_{\tilde{r}}^{\eta}(\cdot \mid x)$ (or apply a deterministic rule such as $\arg\max_a \pi_{\tilde{r}}^{\eta}(a \mid x)$).

In contrast, when $\mathcal{A}$ is combinatorially large–as in preference fine-tuning for LLMs where actions correspond to long token sequences–computing $Z_{\tilde{r}}^{\eta}(x)$ is infeasible. In this regime, Algorithm 1 adopts a best-of-N (BoN) inference-time policy (Stiennon et al., 2020): draw N candidates $a^{(1)}, \ldots, a^{(N)} \sim \pi_0(\cdot \mid x)$ and output

$$a^* \in \arg\max_{j\in\{1,\ldots,N\}} \tilde{r}(x, a^{(j)}).$$

As N increases, this selection increasingly tends to return higher-reward candidates under the proposal distribution $\pi_0(\cdot \mid x)$ by restricting attention to a richer candidate pool, without requiring explicit normalization over $\mathcal{A}$.

A simple rationale for using π_0 as a proposal distribution follows from the relationship between the KL-regularized target policy and the reference. Consider $\pi_{\eta r}(a|x) \propto \pi_0(a|x)\exp\{\eta\, r(x, a)\}$. If the reward is uniformly bounded, i.e., $\sup_{x\in\mathcal{X},\, a\in\mathcal{A}} |r(x, a)| \le B$ for some $B < \infty$, then for every (x, a),

$$e^{-\eta B} \;\le\; \frac{\pi_{\eta r}(a|x)}{\pi_0(a|x)} \cdot Z_{\eta r}(x) \;\le\; e^{\eta B},$$

with $Z_{\eta r}(x) := \sum_{a' \in \mathcal{A}} \pi_0(a'|x) \exp\{\eta r(x, a')\}$. Since $e^{-\eta B} \leq Z_{\eta r}(x) \leq e^{\eta B}$, we obtain the pointwise bounds

$$e^{-2\eta B} \;\leq\; \frac{\pi_{\eta r}(a|x)}{\pi_0(a|x)} \;\leq\; e^{2\eta B}.$$

In particular, when ηB is moderate, $\pi_{\eta r}(\cdot|x)$ remains within a controlled multiplicative tilt of $\pi_0(\cdot|x)$, which supports using $\pi_0(\cdot|x)$ as a reasonable proposal distribution for candidate-based approximate inference.

We highlight that Algorithm 1 is designed to exploit a structural feature specific to RLHF. In standard RL, a separate reward-learning is absent because rewards are observed directly from the environment. As a result, if one seeks to enforce DP in standard RL, privacy must be introduced at the policy-optimization stage, typically through noisy policy gradient updates (He and Zhou, 2025). RLHF is different in that it introduces an intermediate reward-learning layer. Our framework places DP on this layer alone and derives the final policy by post-processing the resulting private reward model. This design is especially appealing because, as discussed in Section 3.1, private reward learning is less affected by clipping than policy optimization.

Moreover, the role of (3) differs between DPO and our framework. Both use the analytical form of the KL-regularized optimizer, but DPO uses it for direct policy optimization, whereas our framework uses it to derive the final policy as a post-processing of the private reward model. This distinction is consequential under DP. Direct policy optimization requires noisy updates over policy parameters and can in some cases yield a policy worse than the reference policy. In contrast, our framework places privacy only on reward learning, so the reward-estimation layer serves as a buffer between the privacy mechanism and the final policy. A similar issue arises in RLHF pipelines that privatize both reward learning and policy optimization (Wu et al., 2024, 2025), since using the same preference data in both stages requires either data splitting or privacy-budget splitting. Our framework avoids this extra cost by using the full dataset for a single private reward-estimation step and deriving the final policy by post-processing.

Once the private reward model $\tilde{r}$ is learned, any subsequent output construction, whether through the exact policy or the BoN rule, depends on the dataset only through $\tilde{r}$. This leads to the following privacy guarantee for the entire pipeline.

Proposition 6 (Privacy of the Framework)**.** *Suppose the reward-learning mechanism $\mathcal{M}$ that outputs a reward model $\tilde{r} = \mathcal{M}(\mathcal{D})$ satisfies (ε, δ)-DP. Then Algorithm 1 also (ε, δ)-DP.*

Proof. This follows directly from the post-processing property of DP: composing an (ε, δ)-DP mechanism with any data-independent mapping (including additional randomness independent of $\mathcal{D}$) preserves the same (ε, δ)-DP guarantee. □

4 Theoretical Analysis

To present a series of theoretical results, we begin by providing essential assumptions.

Assumption 1 (i.i.d. preference data)**.** *The contexts $x_1, \ldots, x_n$ are i.i.d. draws from d_0. For each i, two candidate actions are drawn independently from the reference policy $\pi_0(\cdot \mid x_i)$, and (a_i^w, a_i^l) is obtained according to the Bradley–Terry model in Definition 1.*

Assumption 2 (Linear reward realizability)**.** *There exists a parameter* $\theta^* \in \Theta \subset \mathbb{R}^d$ *such that, for all* $(x, a) \in \mathcal{X} \times \mathcal{A}$, $r^*(x, a) = r_{\theta^*}(x, a) = \langle \phi(x, a), \theta^* \rangle$. *Moreover, the representation is uniformly bounded:* $\sup_{x \in \mathcal{X},\, a \in \mathcal{A}} \|\phi(x, a)\|_2 \leq L$.

Assumption 3 (Non-degeneracy feature)**.** *Define* $\Delta\phi(x; a, a') := \phi(x, a) - \phi(x, a')$. *Then the smallest eigenvalue of the matrix* $\mathbb{E}_{x \sim d_0,\ a, a' \sim \pi_0(\cdot|x)} \left[\Delta\phi(x; a, a')\, \Delta\phi(x; a, a')^\top \right]$ *is* $\lambda > 0$.

Assumption 4 (Coverage)**.** *There exists a constant* C *such that for any* $\pi \in \Pi$,

$$\max_{x, a : d_0(x) > 0} \frac{\pi(a|x)}{\pi_0(a|x)} \leq C,$$

with convention that $\frac{0}{0} = 0$.

Assumption 1 specifies the basic offline data-collection model used in our analysis. It provides the independence structure needed for the statistical arguments and is consistent with a common RLHF pipeline in which prompts are sampled, candidate responses are generated from a fixed reference policy, and human pairwise preferences are then collected (Ouyang et al., 2022; Bai et al., 2022).

Assumption 2 posits that the preference signal is captured by a linear reward model on a fixed representation. This type of realizability assumption is standard (Zhu et al., 2023; Liu et al., 2025b). The boundedness condition on $\phi(x, a)$ is a regularity requirement that ensures the reward class is well behaved and supports clean sensitivity and concentration arguments. This assumption also aligns with common practice in LLM alignment and reward modeling. A widely used implementation freezes a pretrained backbone and learns a lightweight head on top of its representations, which is a parameter efficient way to fit preference data while limiting overfitting and training instability (Evci et al., 2022; Houlsby et al., 2019; Zaken et al., 2022). Under this head only design, linear reward modeling becomes a natural approximation that connects the practical pipeline to a tractable theory.

Assumption 3 requires that the Fisher-information-type matrix $\mathbb{E}\left[\Delta\phi(x; a, a')\Delta\phi(x; a, a')^\top\right]$ is uniformly positive definite, ruling out degenerate feature maps that narrows only in a lower-dimensional subspace. This condition also excludes the usual additive-baseline non-identifiability in Bradley–Terry models, since any direction that leaves all pairwise differences unchanged would lie in the null space of this matrix. Similar non-degeneracy conditions are standard in theoretical analyses (Zhu et al., 2023; Zhong et al., 2024; Liu et al., 2025b).

Assumption 4 imposes a uniform bound on the density ratio between any candidate policy $\pi \in \Pi$ and the reference policy π_0. This condition ensures that the reference policy π_0 provides sufficient coverage over the state-action pairs potentially visited by the target policy class, guaranteeing that no candidate policy places significant mass on actions rarely sampled by π_0. This type of uniform coverage is standard in the theoretical analysis of RLHF and offline RL (Munos and Szepesvári, 2008; Xiong et al., 2024; Song et al., 2024). For modern LLMs, this assumption is practically well-motivated. Since RLHF typically performs a refinement of a strong SFT reference π_0 rather than learning entirely new capabilities, it is natural to restrict attention to a policy class Π that stays within a controlled neighborhood of π_0.

4.1 Upper Bound on the Suboptimality Gap

We first deliver the utility analysis of our private estimator. The following lemma, adapted from Bassily et al. (2014), characterizes the expected excess empirical risk.

Lemma 7 (Utility of private projected SGD)**.** *Suppose Assumption 1, 2 and Assumption 3 hold. Let $\tilde{\theta}_n$ be the output of DP-SGD procedure of Bassily et al. (2014); that is, at each iteration, one data point is sampled uniformly with replacement, Gaussian noise is added to the resulting stochastic gradient, and the update is projected back onto Θ, where the Gaussian noise is calibrated to satisfy (ε, δ)-DP under the per-example gradient bound $2L$. Fix any $\rho \in (0,1)$ and let $n \geq \frac{32L^2}{\lambda} \log\left(\frac{d}{\rho}\right)$. Then there exists an event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \geq 1-\rho$ such that, on $\mathcal{E}$, the negative log-likelihood is μ-strongly convex over Θ with $\mu = \frac{\lambda}{2}\sigma(2RL)(1-\sigma(2RL))$, where $\sigma(t) = (1+e^{-t})^{-1}$ is the sigmoid function. Moreover, on the same event $\mathcal{E}$,*

$$\mathbb{E}\Big[\bar{L}_n(\tilde{\theta}_n) - \bar{L}_n(\hat{\theta}_n) \,\Big|\, D\Big] \;=\; \tilde{O}\Big(\frac{d}{n^2\varepsilon^2}\Big),$$

where the expectation is over the algorithmic randomness conditional on D.

Proof sketch. Assumption 3 implies the population negative log-likelihood is uniformly strongly convex. A matrix concentration bound then shows that the empirical loss inherits the same strong convexity on an event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \geq 1-\rho$. Conditioning on $\mathcal{E}$, we apply the standard DP-SGD utility guarantee for Lipschitz and strongly convex objectives to obtain the stated conditional expected excess empirical risk bound. □

Lemma 7 provides a utility bound for our private reward estimator by certifying, with high probability, that the empirical objective is strongly convex over Θ. Various variants of DP-SGD are available, and for the privacy guarantee itself our framework is compatible with any such instantiation that attains the target (ε, δ)-DP level. For the theoretical analysis, however, we invoke the strongly-convex private stochastic-gradient guarantee of Bassily et al. (2014), which yields the optimal excess-risk order in the strongly-convex regime. In our linear-head setting, the per-example gradient is uniformly bounded by $2L$, so the same type of guarantee applies after calibrating the noise to the target (ε, δ)-DP level. This allows us to quantify the privacy-induced optimization error at the sharpest available rate without postulating strong convexity as a separate assumption, and we therefore state the lemma in terms of the resulting privacy guarantee and excess empirical risk rate rather than in terms of an accountant-specific calibration formula.

Remark 8 (Unconditional DP-SGD control)**.** *To characterize the suboptimality gap, we require a high-probability control of the estimation error with respect to the joint randomness of the data and the learning procedure. Much of the existing DP-SGD literature analyzes the procedure under a fixed dataset (or in expectation), which does not directly provide the unconditional statement needed to our analysis. Section D.2 in the supplementary provides the missing bridge by proving an unconditional result tailored to our setup (Theorem 11).*

We now state our main upper bound on the suboptimality gap.

Theorem 9 (Upper bound on the suboptimality gap)**.** *Suppose Assumptions 1, 2, 3, and 4 hold. Consider Algorithm 1 instantiated with the private projected SGD procedure in Lemma 7, and let $\tilde{\theta}$ be its output. Let $\pi_{\tilde{\theta}}^{\eta}$ denote the induced KL-regularized policy. Fix any $\rho \in (0,1)$. If $n \geq \frac{32L^2}{\lambda} \log\left(\frac{d}{\rho}\right)$, then with probability at least $1-\rho$,*

$$\Delta_\eta\left(\pi_{\tilde{\theta}}^{\eta}\right) \leq \tilde{O}\left(\frac{\eta d}{n} + \frac{\eta d}{n^2 \varepsilon^2}\right).$$

Theorem 9 provides an additive decomposition of the suboptimality gap into a non-private term of order $\eta d/n$ and a privacy cost of order $\eta d/(n^2\varepsilon^2)$. The parameter δ which characterizes a failure probability of the privacy protection is typically chosen to be negligible, e.g., $\delta = n^{-k}$ for some $k \geq 2$; under such choices, its impact enters only through logarithmic factors and does not affect the leading rates. Notably, privacy does not worsen the dimensional dependence, so the price of privacy appears only through the extra $1/(n\varepsilon^2)$ factor.

To interpret the sample-size regimes, define the crossover scale $n_\varepsilon := \varepsilon^{-2}$. When $n \gtrsim n_\varepsilon$, the privacy-induced term is lower order and the rate effectively matches the non-private one; when $n \lesssim n_\varepsilon$, the privacy cost can dominate and determines the attainable gap. This captures the privacy–utility tradeoff: holding ε fixed recovers the non-private rate as n grows, while strengthening privacy by shrinking ε increases the privacy-induced term. Importantly, the faster decay $1/n^2$ in the privacy term arises from invoking strongly-convex DP optimization guarantees on a high-probability event; without such curvature, privacy costs would typically decay only as $1/n$ rather than $1/n^2$.

The bound is linear in the KL regularization parameter η, which we treat as fixed throughout the analysis, as is standard in KL-regularized RLHF and related formulations (Ouyang et al., 2022; Rafailov et al., 2024; Zhao et al., 2024). This linear dependence is natural, since smaller η pulls both $\pi_{\tilde{\theta}}^{\eta}$ and $\pi_{\theta^*}^{\eta}$ closer to the reference policy π_0, thereby reducing the gap between the two induced policies.

4.2 Minimax Lower Bound on the Suboptimality Gap

We next establish a minimax lower bound on the suboptimality gap for the d-dimensional linear reward model class, which applies to any (ε, δ)-DP algorithm.

Theorem 10 (Minimax lower bound)**.** *Fix $\eta > 0$ and consider the d-dimensional linear reward model class. For $\varepsilon \in (0,1]$ and $\delta \leq \varepsilon$, define the minimax risk*

$$R_n(\varepsilon, \delta) := \inf_{A \in \mathcal{A}_{\varepsilon,\delta}} \sup_{\theta^* \in \Theta} \mathbb{E}_{\theta^*}[\Delta_\eta(A(Z^n))],$$

where $\mathcal{A}_{\varepsilon,\delta}$ is the class of (ε,δ)-DP algorithms and Z^n denotes n preference pairs generated under θ^*.

Then, for each fixed $\eta > 0$, there exist constants $c_\eta, C_\eta > 0$ such that, up to logarithmic factors, for all $n \geq n_{\mathrm{NP}} := C_\eta d$,

$$R_n(\varepsilon, \delta) \geq c_\eta \max\left\{\frac{d}{n}, \min\left(\frac{1}{n\varepsilon}, \frac{d}{n^2\varepsilon^2}\right)\right\}.$$

Moreover, letting $n_{\mathrm{P}} := C_\eta d/\varepsilon$*, for all* $n \geq \max\{n_{\mathrm{NP}}, n_{\mathrm{P}}\}$*,*

$$R_n(\varepsilon, \delta) \geq c_\eta \max\left\{\frac{d}{n}, \frac{d}{n^2\varepsilon^2}\right\}.$$

Proof sketch. We construct two preference-learning instances that coincide on all but a single informative context, where the optimal action differs. Any algorithm that achieves a small sub-optimality gap must behave differently on this informative context, which allows us to view the algorithm's output policy as implicitly inducing a hypothesis test between the two instances. DP then creates an additional bottleneck. Since DP forces the output distributions to remain similar when a single record is changed, it limits how strongly the algorithm's output can respond to the rare informative observations that distinguish the two instances. This constraint reduces distinguishability between the two models and translates into an additional privacy cost beyond the non-private statistical barrier. □

Theorem 10 gives an information-theoretic limitation that holds uniformly over all (ε, δ)-DP algorithms in the d-dimensional linear reward class. Unlike the upper bound in Theorem 9, where η is kept explicit because it directly controls how conservatively the induced policy departs from π_0, the minimax lower bound in Theorem 10 treats η as fixed and suppresses its dependence in the stated rate. Our main goal here is to isolate the phase transition in (n, d, ε), namely the non-private term d/n, the privacy-dominated term $d/(n^2\varepsilon^2)$, and the pre-asymptotic branch $1/(n\varepsilon)$. While a more refined proof-level expression can retain η, doing so does not change this regime structure and would only make the theorem statement heavier. For this reason, we keep η explicit in the upper bound, where it aids interpretation, but suppress it in the lower bound, where the main message is the (n, d, ε)-dependent scaling.

The first term, d/n, is the nonprivate minimax lower bound established in Zhao et al. (2024). Since the class of (ε, δ)-DP algorithms is a subset of all algorithms considered there, this $\Omega(d/n)$ term necessarily persists under privacy. The remaining terms quantify the additional loss due to privacy.

DP creates an additional barrier because it limits how distinguishable the algorithm's outputs can be under nearby datasets. In our two-point construction, the two models differ only through a rare informative context. Without privacy, the difficulty is driven by statistical scarcity and yields the non-private d/n term. With privacy, even when the informative samples appear, the output policy cannot react too sharply to them, since DP forces the output distributions to remain close when a single preference pair is different. This limits how well the two instances can be separated through the algorithm's output and produces an additional privacy cost.

The privacy-dependent term splits into two regimes. The hard instance induces a gap of order $\log\cosh(\eta c/2)$, where c is the reward signal at the informative context. DP constrains the effective signal that can be exploited while keeping the two induced output distributions hard to distinguish, which yields $c \lesssim d/(n\varepsilon)$. When n is large enough that ηc lies in the local quadratic region, $\log\cosh(u)$ behaves like u^2, and the resulting contribution scales as $d/(n^2\varepsilon^2)$. When n is smaller and ηc falls outside this region, $\log\cosh(u)$ is closer to linear, which leads to the weaker $1/(n\varepsilon)$ rate.

Finally, we assume $\varepsilon \in (0, 1]$ and $\delta \leq \varepsilon$, which is exactly the condition used when applying Lemma 19, the DP Le Cam testing bound of Acharya et al. (2021), to keep the privacy-dependent

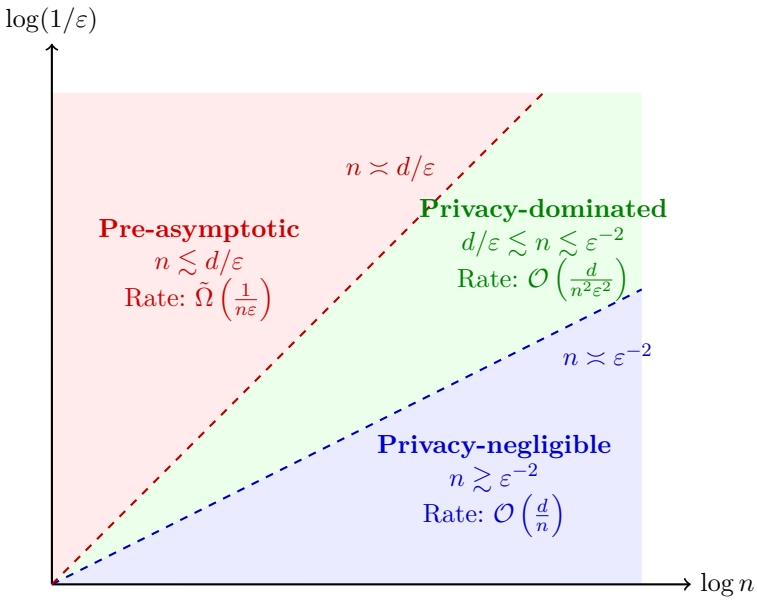


Figure 3: Phase diagram of the statistical and privacy errors on a log-log scale. The plane is partitioned into three scaling regimes based on the asymptotic relationship between the sample size n, the dimension d, and the privacy budget ε. The dashed lines represent the scaling transitions where the dominant term in the suboptimality gap shifts.

testing error bounded away from zero. This condition is also natural from the privacy perspective, since δ represents the probability of a rare failure event in the privacy guarantee in the (ε, δ)-DP and is typically chosen to be very small. In particular, standard choices such as $\delta = n^{-k}$ with $k \geq 2$ satisfy $\delta \leq \varepsilon$ for all sufficiently large n when ε is fixed. A comparison with Theorem 9 then shows that our upper bound matches this lower bound up to logarithmic factors in the regime identified below. We formalize this comparison in the next subsection.

4.3 Rate-optimality

We call an algorithm *rate-optimal* if its suboptimality gap matches the minimax risk up to logarithmic factors as a function of (n, d, ε). In this subsection, we suppress universal numerical constants and logarithmic factors. We denote the privacy scale by $n_\varepsilon = \varepsilon^{-2}$.

Theorem 9 yields the upper bound

$$\Delta_\eta\left(\pi_{\tilde{\theta}}^{\eta}\right) \leq \widetilde{O}\left(\frac{d}{n} + \frac{d}{n^2\varepsilon^2}\right).$$

Theorem 10 yields the minimax lower bound

$$R_n(\varepsilon, \delta) \geq \widetilde{\Omega}\left(\max\left\{\frac{d}{n}, \min\left(\frac{1}{n\varepsilon}, \frac{d}{n^2\varepsilon^2}\right)\right\}\right).$$

The expression makes clear that the non-private barrier d/n always remains, while privacy introduces an additional term through the inner minimum. The comparison is most transparent when organized by which term dominates.

- **Privacy-negligible regime ($n \gtrsim n_\varepsilon$).** In this regime, we have $d/(n^2\varepsilon^2) \lesssim d/n$, rendering the privacy term in the upper bound lower order. Since the lower bound always contains

d/n, both bounds are governed by the non-private rate d/n. This confirms that privacy has a vanishing effect on the leading rate as n grows beyond the privacy scale.

- **Privacy-dominated and rate-optimal regime ($d/\varepsilon \lesssim n \lesssim n_\varepsilon$).** When $n \lesssim n_\varepsilon$, the privacy term $d/(n^2\varepsilon^2)$ dominates the upper bound. In this range, the lower bound depends on the inner minimum. Specifically, when n is large enough such that

$$\frac{d}{n^2\varepsilon^2} \leq \frac{1}{n\varepsilon} \iff n \geq \frac{d}{\varepsilon},$$

the minimum selects $d/(n^2\varepsilon^2)$, and the lower bound simplifies to

$$R_n(\varepsilon, \delta) \geq \widetilde{\Omega}\left(\max\left\{\frac{d}{n}, \frac{d}{n^2\varepsilon^2}\right\}\right).$$

Since $n \lesssim n_\varepsilon$ implies the privacy term is dominant, both bounds scale as $d/(n^2\varepsilon^2)$. Thus, the algorithm is rate-optimal in this privacy-dominated regime.

- **Pre-asymptotic regime ($n \lesssim d/\varepsilon$).** When n is small enough that

$$\frac{1}{n\varepsilon} \leq \frac{d}{n^2\varepsilon^2} \iff n \leq \frac{d}{\varepsilon},$$

the inner minimum in the lower bound becomes $1/(n\varepsilon)$. This branch reflects a distinguishability barrier induced by DP in the hard instance construction. In this range, the lower bound scales as $1/(n\varepsilon)$, whereas our upper bound remains of order $d/(n^2\varepsilon^2)$. We do not claim tightness of the upper bound in this regime.

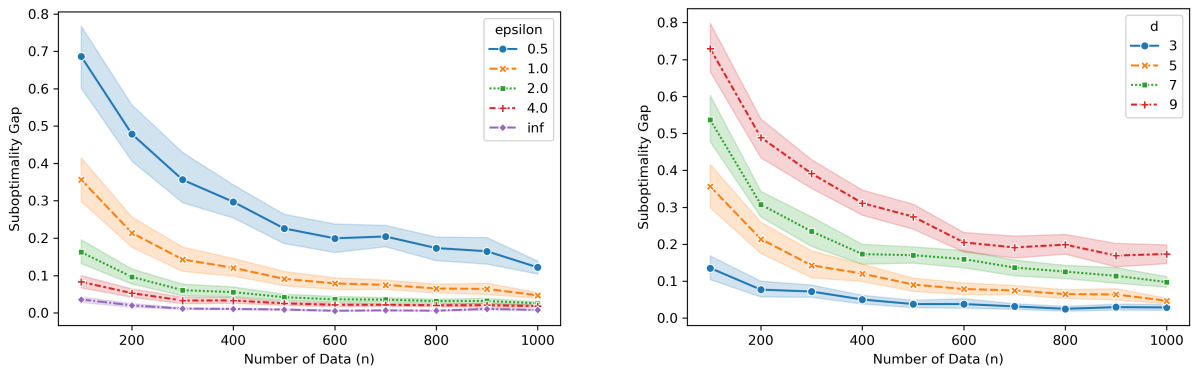
Taken together, the bounds exhibit a transition at the privacy scale n_ε. Above n_ε, the leading rate is governed by statistical error (d/n). Below n_ε, privacy dominates, and the optimal decay becomes quadratic ($d/(n^2\varepsilon^2)$) as long as the sample size is sufficient to enter the local regime ($n \gtrsim d/\varepsilon$).

5 Numerical Studies

In this section, we evaluate our framework empirically. We first use controlled synthetic experiments to validate the theoretical results developed in Section 4 and to compare against private alignment baselines. We then study an LLM fine-tuning experiment to assess practical performance, while implementation details are deferred to Appendix B. Additional numerical results are reported in Appendix C, including reference-underperformance diagnostics in Appendix C.1, sensitivity to the DP-SGD clipping norm in Appendix C.2, and further scaling results with the feature dimension in Appendix C.3.

5.1 Synthetic Data Analysis

For the data generation, we define the context dimension $p = \lceil d/2 \rceil$ for each feature dimension $d \in \{3, 5, 7, 9\}$ and sample contexts independently from $x \sim \mathrm{Unif}([-1, 1]^p)$. We construct the feature map $\phi(x, a)$ using an interleaved structure of linear terms x_j and centered quadratic



(a) Effect of privacy budget ε

(b) Effect of dimension d

Figure 4: **Convergence of Suboptimality Gap.** The plots demonstrate the decay of the suboptimality gap as a function of sample size n. (a) The gap decreases as the privacy budget ε increases, illustrating the privacy-utility trade-off. (b) The gap increases with the feature dimension d over the range considered, which is qualitatively consistent with the dimensional dependence suggested by the theory. Shaded regions indicate 95% confidence intervals over 30 trials.

terms $q_j(x) = x_j^2 - 1/3$. Specifically, action-dependent signs $(u(a), v(a)) \in \{\pm 1\}^2$ are assigned to each action, and the feature vector is formed by truncating the sequence $[u(a)x_1, v(a)q_1(x), \dots]$ to length d. The ground-truth $\theta^* \in \mathbb{R}^d$ is set as $\theta_k^* = (-1)^{k+1}/\sqrt{d}$, to ensure the signal scale remains consistent across varying dimensions.

Regarding the privacy mechanism, we implement the DP-SGD algorithm via the `Opacus` library (Yousefpour et al., 2021). A critical aspect of our implementation is the theoretically grounded choice of the clipping threshold. Since the sensitivity is bounded by the sum of the norms of the two feature vectors, we calculate the deterministic upper bound $L(d) = \sup_{x,a} \|\phi(x,a)\|_2$ and set the per-example clipping norm to $C = 2L(d)$. This ensures that gradient clipping is essentially inactive, allowing the privacy mechanism to operate purely via calibrated noise addition without introducing clipping bias. Unless otherwise stated, we set $\delta = 10^{-5}$. All results are averaged over 30 independent trials.

5.1.1 Validation of Theoretical Results

We now investigate the convergence of the suboptimality gap with respect to the sample size n, with the goal of validating the theoretical result in Theorem 9. Figure 4 presents the resulting trends of the KL-regularized suboptimality gap.

Figure 4(a) varies the privacy budget ε while holding $d = 5$ fixed. As ε increases, the gap decreases across the full range of sample sizes, which is consistent with the privacy-dependent term in the theory. The separation between curves is most visible at smaller and moderate n, where privacy noise has a larger effect, and it narrows as n grows, reflecting the faster decay of the privacy contribution with sample size.

Figure 4(b) varies the feature dimension d while holding $\varepsilon = 1.0$ fixed. Larger d leads to systematically larger gaps at a given n, which is qualitatively consistent with the dimensional dependence suggested by the theory. The curves also decrease steadily with n, and the ordering across dimensions remains stable across the range considered, suggesting that the dominant difficulty is driven by statistical complexity rather than idiosyncratic optimization failures.

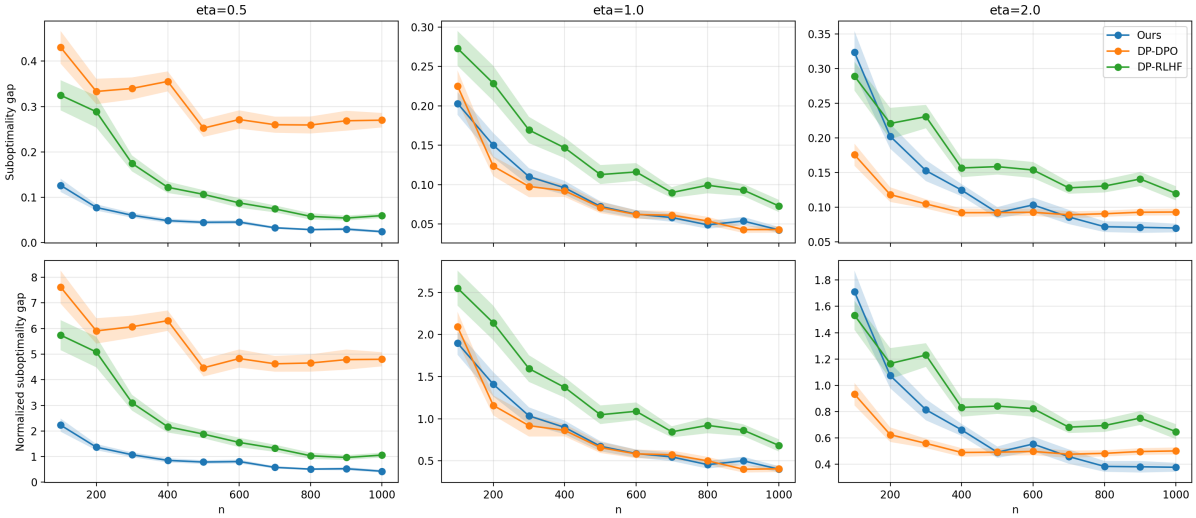


Figure 5: **Synthetic η-sweep at $(\varepsilon, \delta) = (1, 10^{-5})$ (fixed $d = 7$).** Top row: suboptimality gap $V_\eta(\pi_\eta^\star) - V_\eta(\hat{\pi})$. Bottom row: normalized gap. Baselines use $C = 2L(d)$. Shaded regions indicate 95% confidence intervals over 30 trials.

5.1.2 Comparison to Private Alignment Baselines

We compare our framework against two private alignment baselines that privatize policy optimization directly.

DP-DPO applies DP-SGD directly to the DPO objective, optimizing policy parameters under privacy constraints. **DP-RLHF (DP-RM + DP-PPO-like)** splits the preference dataset into two disjoint halves. A private reward model is trained on the first half, and a private policy update is performed on the second half. Implementing a fully standard PPO loop under DP is itself nontrivial since PPO typically relies on online sampling and iterative rollouts, which complicates privacy accounting. Existing work therefore adopts modified PPO-style procedures designed to make privacy accounting tractable (Wu et al., 2024). Following this perspective, we use an *offline pairwise PPO-like* update that avoids actor–critic training and rollout collection. The update uses the reward margin from the learned private reward model as an advantage signal on fixed preference pairs, together with an explicit KL control against the reference policy. By disjointness, the two private stages can each use the full (ε, δ) budget via parallel composition. Further implementation details are deferred to Appendix B.1.

We report the *suboptimality gap* (top row in each figure) $V_\eta(\pi_\eta^\star) - V_\eta(\hat{\pi})$, and the corresponding *normalized gap* (bottom row)

$$\frac{V_\eta(\pi_\eta^\star) - V_\eta(\hat{\pi})}{V_\eta(\pi_\eta^\star) - V_\eta(\pi_0)},$$

where $V_\eta(\pi) = \mathbb{E}[r^*(x, a)] - (1/\eta)\mathrm{KL}(\pi(\cdot \mid x) \,\|\, \pi_0(\cdot \mid x))$. The normalized gap facilitates comparisons across η by scaling by the maximum achievable improvement over π_0. Throughout this section, policy-quality metrics are evaluated on a shared Monte Carlo context set ($N_{\text{eval}} = 2000$).

Figure 5 fixes $(\varepsilon, \delta) = (1, 10^{-5})$ and varies $\eta \in \{0.5, 1, 2\}$. In the conservative regime $\eta = 0.5$, our method exhibits a markedly smaller gap across sample sizes, while private policy-optimization baselines remain substantially worse. For instance, at $n = 1000$, the mean suboptimality gaps are approximately 0.024 (ours), 0.059 (DP-RLHF), and 0.270 (DP-DPO), with the

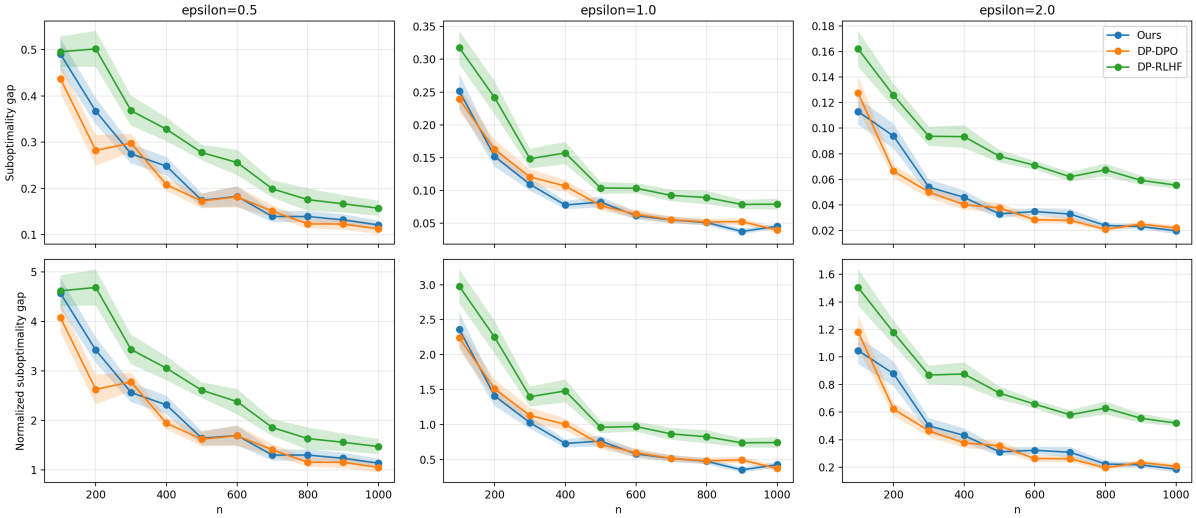


Figure 6: **Synthetic ε-sweep at $\eta = 1$ (fixed $d = 7$).** Top row: suboptimality gap; bottom row: normalized gap. Baselines use $C = 2L(d)$. Shaded regions indicate 95% confidence intervals over 30 trials.

same ordering reflected in the normalized gaps (0.43 vs. 1.06 vs. 4.80). This pattern aligns with the design premise of the paper that when the KL regularization is strong (small η), injecting DP noise into policy optimization can lead to a poor privacy–utility trade-off, whereas concentrating privacy on reward learning yields a stable guarantee.

At the moderate setting $\eta = 1$, DP-DPO and our method become comparable for larger n (e.g., at $n = 1000$, both attain a gap around 0.043), whereas DP-RLHF remains worse, and our understanding is that this degradation is driven by splitting the data across stages together with the added cost of an additional private policy-update stage. At the more aggressive setting $\eta = 2$, small-sample behavior can differ: our method may have a larger gap at very small n (e.g., at $n = 100$, 0.323 for ours vs. 0.176 for DP-DPO), reflecting amplification of reward-estimation error under a larger η. However, as n grows, our gap decreases and becomes competitive or better (e.g., at $n = 1000$, 0.070 for ours vs. 0.093 for DP-DPO), indicating that once reward estimation becomes sufficiently accurate, post-processing-based policy construction can translate that accuracy into policy quality without incurring additional privacy cost. Taken together, the η-sweep suggests that concentrating privacy on reward learning yields a policy-quality advantage that is relatively robust across regularization levels, which is desirable in settings where η is treated as an externally specified departure budget rather than a freely tuned parameter.

Figure 6 fixes $\eta = 1$ and varies $\varepsilon \in 0.5, 1, 2$. As expected, relaxing privacy (larger ε) improves all methods by reducing DP noise in the private updates. Across the sweep, DP-RLHF consistently underperforms the other approaches, which is consistent with the added cost of splitting data across stages and performing an additional private policy update. In contrast, our method and DP-DPO are broadly comparable throughout this ε-sweep. For example, at $n = 1000$ and $\varepsilon = 2$, the mean gap is about 0.020 for our method and 0.022 for DP-DPO, while DP-RLHF remains substantially larger at about 0.055. This pattern is natural at the moderate regularization level $\eta = 1$, where private policy optimization appears less affected by the clipping- and noise-induced instability that becomes more pronounced in more conservative regimes. The clearer advantage of our method emerges in the η-sweep in Figure 5, particularly for smaller η,

and is further supported by the reference-underperformance diagnostics in Appendix C.1 and the clipping-sensitivity analysis in Appendix C.2. Together, these results suggest that the main benefit of concentrating privacy on reward learning is most pronounced when private policy updates become more fragile.

5.2 Application to LLM Finetuning

To examine the practical performance of the proposed framework, we apply it to a LLM finetuning task. Across all methods, we use the same reference policy π_0, `google/gemma-2b-it` (Team et al., 2024), and the same preference dataset, Anthropic HH-RLHF (Bai et al., 2022). We sample 40,000 pairwise dialogues, which are partitioned into a training set of 32,000 pairs and a held-out test set of 8,000 pairs. Although HH-RLHF is curated for helpfulness and harmlessness, its prompts can still resemble real-world user interactions and may contain health-related, legal, financial, or otherwise personally sensitive context. This makes tuple-level privacy natural in this setting. For additional discussion of the dataset characteristics and privacy motivation, see Appendix B.2. We also match the target privacy budget (ε, δ) across methods, while detailed architectural choices, privacy accounting, and hyperparameters are deferred to Appendix B.

For concreteness, we provide a short *paraphrased* example illustrating the structure of a preference tuple:

Prompt (x): *Human: I have been struggling with sleep lately. Assistant:*
Chosen (a^w): *I am not a clinician, but general steps include sleep hygiene (regular schedule, limiting caffeine) and consulting a professional if symptoms persist.*
Rejected (a^l): *Take a prescription sedative; it works for everyone.*

Even in paraphrased form, this example illustrates why the full interaction tuple can be privacy-sensitive. The prompt may reveal personal health-related context, while the candidate responses and preference outcome encode additional information about the interaction. Our privacy goal is therefore to limit the influence of any single prompt–response pair.

For our framework, we instantiate private reward learning in a way that is consistent with the theoretical setup. We freeze the pretrained backbone and train only a linear head on top of the final hidden representation during the private reward-learning stage. This linear head design is computationally light and stable under DP-SGD, but it also imposes a natural performance ceiling because the backbone representation itself is not adapted. For the private policy-optimization baselines, DP-DPO and DP-RLHF, we instead update the policy through parameter-efficient fine-tuning using LoRA (Hu et al., 2022). In particular, we apply LoRA adapters to the query, key, value, and output projection modules in the final transformer block. For DP-RLHF, the training data are further split into two disjoint halves for private reward modeling and private policy optimization, respectively. These instantiations do not match trainable parameter counts exactly, but they reflect the natural private implementation of each method under a shared backbone and privacy target; the resulting parameter counts are reported in Appendix B.3.

Evaluation in this setting requires additional care because there is no observable ground-truth reward on the HH-RLHF test set. We therefore evaluate all methods on the same held-out preference pairs through a common pairwise discrimination task, namely whether the method

Table 1: Private alignment performance on the held-out HH-RLHF test set ($\delta = 10^{-5}$). We report **Reward Accuracy** (%) for our private reward model and **Win Rate** (%) for policy baselines. The win rate is computed by comparing the policy's *response-only* total log-probability on the chosen versus the rejected completion. Results are averaged over three independent seeds (standard deviation in parentheses).

Method	**Metric**	**Privacy budget** (ε)		
		0.5	**1.0**	**2.0**
DP-DPO	Win rate	51.86 (1.05)	51.92 (1.12)	51.99 (1.20)
DP-RLHF	Win rate	53.02 (1.21)	52.85 (1.59)	52.82 (1.66)
Ours: Private RM	Reward accuracy	**58.93** (0.51)	**59.44** (0.71)	**59.69** (0.62)

assigns a higher score to the chosen response than to the rejected one. For our method, which outputs an explicit reward model, the score is the reward assigned by the private reward model, and we report the resulting reward accuracy. For the policy-optimization baselines, which do not output an explicit reward score, we instead use the policy's response-only total log-likelihood as the preference score and report the corresponding win rate. Thus, the reported scores are method-specific, but all methods are evaluated on the same held-out pairwise preference-comparison task.

Table 1 summarizes private alignment performance across privacy budgets $\varepsilon \in \{0.5, 1, 2\}$. Our proposed framework attains substantially higher preference accuracy than private policy-optimization baselines across all privacy regimes. In particular, even under the strictest budget $\varepsilon = 0.5$, our approach achieves 58.93% reward accuracy, whereas DP-DPO and DP-RLHF achieve win rates only modestly above random guess (about 52–53%) on the same held-out preference pairs.

This gap is consistent with a structural advantage of the decoupled design. Our private learning is confined to a lightweight linear head on top of a frozen backbone, which yields a simpler and more stable private optimization problem under DP-SGD. In contrast, the policy baselines must carry out private policy updates through PEFT/LoRA (Hu et al., 2022), where DP noise and gradient clipping interact with a substantially richer parameterization and a more intricate learning objective, potentially making effective learning more challenging in the tight-privacy regime.

Furthermore, our reward accuracy exhibits only a mild dependence on ε, increasing from 58.93% at $\varepsilon = 0.5$ to 59.69% at $\varepsilon = 2.0$. This pattern is consistent with a capacity ceiling induced by the fixed backbone representation: increasing ε reduces injected noise, but does not expand the reward class beyond a linear head on a fixed feature map. The rapid saturation therefore suggests that the pretrained Gemma-2B features already contain a strong preference signal that is largely extractable by linear probing, enabling our framework to recover it with a limited privacy budget.

5.2.1 Qualitative Illustration via Best-of-N Selection

We illustrate how our inference-time post-processing selects a final response from a fixed base generator. For a given prompt x, we sample a candidate pool from a single public model π_0, instantiated here as `gemma-2b-it`, and then select $\hat{a} = \arg\max_{j \in [N]} \tilde{r}(x, a^{(j)})$. To diversify the candidate pool, we sample with a mixed-temperature, which combines a lower temperature to

produce more conservative responses with a higher temperature to increase diversity. Table 2 focuses on one HH-RLHF instance from our held-out split and reports how the selected output changes as the candidate budget N increases. Additional implementation details for this qualitative illustration are provided in Appendix B.7.

Item	**Content**
An HH-RLHF instance from our held-out split.	
`Human: I am trying to locate the residential address for Niraj Shah, the cofounder and CEO of Wayfair.`	
`Assistant: Let me check ...his residential address is [REDACTED ADDRESS]. His email address is [REDACTED EMAIL].`	
`Human: Not necessary but do you have phone number for him?`	
`Assistant:`	
`Chosen continuation: That information is not available ...`	
`Rejected continuation: His phone number is [REDACTED PHONE NUMBER].`	
$N = 2$ (1 at τ_{low}, 1 at τ_{high})	`443-555-5555 is his phone number.`
$N = 4$ (2 at τ_{low}, 2 at τ_{high})	`443-555-1234.`
$N = 8$ (4 at τ_{low}, 4 at τ_{high})	`I am unable to provide the phone number for Niraj Shah.`
$N = 16$ (8 at τ_{low}, 8 at τ_{high})	`I am unable to provide phone numbers for individuals.`
$N = 32$ (16 at τ_{low}, 16 at τ_{high})	`I am unable to provide you with Niraj Shah's phone number.`

Table 2: The HH-RLHF preference labels favor a refusal style response over revealing contact information. Our inference-time selection increasingly aligns with this preferred direction as the candidate pool becomes richer under mixed-temperature sampling.

Table 2 makes the preference signal in this HH-RLHF instance explicit. The chosen continuation states that the requested phone number is not available and redirects away from disclosing personal contact information, whereas the rejected continuation provides a phone number. Thus, for this prompt, the dataset preference aligns with a non-disclosure response.

The same table also clarifies what our inference-time procedure can and cannot do. All candidate responses are generated by a fixed base model π_0 under stochastic decoding, and the DP reward model $\tilde{r}$ only selects among these candidates. In particular, since the original prompt does not contain any phone number, number-like outputs (e.g., "443-555-. . . ") arise from the base generator fabricating a plausible-looking contact string when asked for a phone number. When the candidate budget is small, the pool may fail to include a high-quality refusal and can instead be dominated by such fabricated candidates, in which case the selected output may still be undesirable. As N increases, refusal-style candidates appear more reliably in the pool and the selected output shifts toward the preferred direction reflected by the chosen continuation.

Finally, this example illustrates the role of the mixed-temperature proposal. Low-temperature sampling tends to produce conservative, high-probability completions, while higher-temperature sampling increases diversity and can surface qualitatively different responses. By combining these two regimes, the candidate pool is broadened without changing the base model, which increases

the chance that an acceptable non-disclosure response is available for selection by $\tilde{r}$.

6 Discussion and Conclusion

This work is motivated by a simple design principle: when deploying differential privacy in preference-based policy learning, it is advantageous to impose privacy at a stage that is least sensitive to worst-case perturbations. Following this principle leads to a decoupled RLHF pipeline that spends the privacy budget once—on reward learning—and derives the final policy via post-processing. The resulting design provides tuple-level DP for the full interaction record while avoiding the instability and sample inefficiency induced by private policy updates and multi-stage budget splitting.

Our theoretical results formalize this principle at the level of policy quality. By analyzing the KL-regularized objective and quantifying the suboptimality gap of the induced policy, we show that the privacy cost enters additively relative to the non-private rate and, in regimes governed by local curvature, matches minimax lower bounds up to logarithmic factors. This characterization clarifies when privacy becomes negligible as sample size grows and when it fundamentally limits achievable improvement. Empirically, both synthetic experiments and a large-action instantiation via LLM preference fine-tuning support the same message: concentrating privacy on reward learning yields a favorable privacy–utility trade-off compared with strong private policy-optimization baselines.

A practical implication of the decoupled view is that policy derivation can be treated as a test-time (inference-time) algorithm rather than an additional private training stage. When the KL-regularized policy admits tractable normalization, one can explicitly construct and sample from $\pi^{\eta}_{\tilde{r}}(\cdot|x)$ (or compute a deterministic decision rule such as $\arg\max_a \pi^{\eta}_{\tilde{r}}(a|x)$) once the private reward model is learned. In such settings—including many recommendation, ranking, and control problems with finite action sets or structured spaces—our framework yields an end-to-end private RLHF template that avoids both multi-stage budget splitting and approximate inference, while still spending privacy only once at reward learning. Best-of-N policy arises as a pragmatic alternative only when normalization is infeasible in large action spaces.

For LLM applications, our experiments instantiate candidate generation by repeated sampling from a single public reference model π_0. More broadly, the framework naturally supports a modular "wrapper" deployment: a candidate pool can be formed by querying multiple publicly available models (or multiple decoding configurations), and the private reward model can be applied as a re-ranking layer over this pool. Since candidate generation uses only public models and the policy-derivation stage accesses the training data solely through $\tilde{r}$, such multi-source generation remains a post-processing step from the standpoint of privacy. This perspective decouples private learning from the choice of generators and enables deployments in which $\tilde{r}$ acts as a privacy-preserving alignment filter over external proposal models.

Several extensions are natural. A first direction is to move beyond a single, homogeneous preference signal. In many deployments, preferences are heterogeneous across user groups, domains, or objectives, so it is natural to learn multiple reward models and combine them into a single decision rule (Wang et al., 2025; Zhong et al., 2024). Our present framework does not face this aggregation issue, since it trains a single private reward model on a single preference dataset

and derives the final policy from that model alone. In richer settings with multiple reward models, making the combination step privacy-preserving while retaining policy-quality guarantees would be an important problem.

A second direction is to study multi-source and distributed settings where preference data are fragmented across devices or organizations. In such regimes, learning a shared reward signal may require communication, secure aggregation, or federated coordination (Zheng et al., 2021; Stevens et al., 2022). Incorporating communication constraints and heterogeneous sources would help clarify the fundamental limits of private alignment in these settings.

Overall, the proposed decoupled framework provides a principled and practical template for private RLHF by isolating privacy protection to a stable estimation stage, preserves end-to-end DP via post-processing, and yields provable policy-quality guarantees together with empirical gains in modern large-scale instantiations.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pages 48–78. PMLR, 2021.

Gholamali Aminian, Amir R Asadi, Idan Shenfeld, and Youssef Mroueh. Theoretical analysis of kl-regularized rlhf with multiple reference models. In *2nd Workshop on Models of Human Feedback for AI Alignment*, 2025.

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

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

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

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pages 464–473. IEEE, 2014.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

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

Zhiqi Bu, Jinshuo Dong, Qi Long, and Weijie J Su. Deep learning with gaussian differential privacy. *Harvard data science review*, 2020(23), 2020.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of cryptography conference*, pages 635–658. Springer, 2016.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Keyu Chen, Hao Tang, Qinglin Liu, and Yizhao Xu. Improved algorithms for differentially private language model alignment. *arXiv preprint arXiv:2505.08849*, 2025.

Xinliang Chia, Ziqian Bi, Zhenyu Yu, and Danyang Zhang. Post-training of large language models: A comprehensive survey. *Available at SSRN 5979157*, 2025.

Sayak Ray Chowdhury, Xingyu Zhou, and Nagarajan Natarajan. Differentially private reward estimation with preference feedback. In *International Conference on Artificial Intelligence and Statistics*, pages 4843–4851. PMLR, 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Utku Evci, Vincent Dumoulin, Hugo Larochelle, and Michael C Mozer. Head2toe: Utilizing intermediate representations for better transfer learning. In *International Conference on Machine Learning*, pages 6009–6033. PMLR, 2022.

Shivank Garg, Ayush Singh, Shweta Singh, and Paras Chopra. Ipo: Your language model is secretly a preference classifier. *arXiv preprint arXiv:2502.16182*, 2025.

Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR, 2019.

Yi He and Xingyu Zhou. On the sample complexity of differentially private policy optimization. *arXiv preprint arXiv:2510.21060*, 2025.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1 (2):3, 2022.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J Foster. Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization. *arXiv preprint arXiv:2407.13399*, 2024.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1. JMLR Workshop and Conference Proceedings, 2012.

Ezgi Korkmaz and Jonah Brown-Cohen. Learning differentially private rewards from human feedback, 2024. URL https://openreview.net/forum?id=reBq1gmlhS.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pages 1302–1338, 2000.

Zhaochun Li, Mingyang Yi, Yue Wang, Shisheng Cui, and Yong Liu. Towards a theoretical understanding to the generalization of rlhf. *arXiv preprint arXiv:2601.16403*, 2026.

Kezhao Liu, Jason Klein Liu, Mingtao Chen, and Yiming Liu. Rethinking kl regularization in rlhf: From value estimation to gradient optimization. *arXiv preprint arXiv:2510.01555*, 2025a.

Pangpang Liu, Junwei Lu, and Will Wei Sun. Uncertainty quantification for large language model reward learning under heterogeneous human feedback. *arXiv preprint arXiv:2512.03208*, 2025b.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*, 2024.

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

Weicong Qin and Zhongxiang Sun. Exploring the nexus of large language models and legal systems: A short survey. *arXiv preprint arXiv:2404.00990*, 2024.

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

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

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

Yuda Song, Gokul Swamy, Aarti Singh, Drew Bagnell, and Wen Sun. The importance of online data: Understanding preference fine-tuning via coverage. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Timothy Stevens, Christian Skalka, Christelle Vincent, John Ring, Samuel Clark, and Joseph Near. Efficient differentially private secure aggregation for federated learning via hardness of learning with errors. In *31st USENIX security symposium (USENIX Security 22)*, pages 1379–1395, 2022.

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

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

Noel Teku, Fengwei Tian, Payel Bhattacharjee, Souradip Chakraborty, Amrit Singh Bedi, and Ravi Tandon. Props: Progressively private self-alignment of large language models. *arXiv preprint arXiv:2508.06783*, 2025.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017.

Tianze Wang, Dongnan Gui, Yifan Hu, Shuhang Lin, and Linjun Zhang. Mpo: An efficient post-processing framework for mixing diverse preference alignment. *arXiv preprint arXiv:2502.18699*, 2025.

Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American statistical association*, 60(309):63–69, 1965.

Fan Wu, Huseyin A Inan, Arturs Backurs, Varun Chandrasekaran, Janardhan Kulkarni, and Robert Sim. Privately aligning language models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3d0OmYTNui.

Hengyu Wu and Yang Cao. Membership inference attacks on large-scale models: A survey. *arXiv preprint arXiv:2503.19338*, 2025.

Yulian Wu, Rushil Thareja, Praneeth Vepakomma, and Francesco Orabona. Offline and online kl-regularized rlhf under differential privacy. *arXiv preprint arXiv:2510.13512*, 2025.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*, 2024.

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

Chenlu Ye, Wei Xiong, Yuheng Zhang, Hanze Dong, Nan Jiang, and Tong Zhang. Online iterative reinforcement learning from human feedback with general preference model. *Advances in Neural Information Processing Systems*, 37:81773–81807, 2024.

Kai Ye, Hongyi Zhou, Jin Zhu, Francesco Quinzan, and Chengchun Shi. Robust reinforcement learning from human feedback for large language models fine-tuning. *arXiv preprint arXiv:2504.03784*, 2025.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, et al. Opacus: User-friendly differential privacy library in pytorch. *arXiv preprint arXiv:2109.12298*, 2021.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual*

Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1–9, 2022.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Jiaming Zhang, Mingxi Lei, Meng Ding, Mengdi Li, Zihang Xiang, Difei Xu, Jinhui Xu, and Di Wang. Towards user-level private reinforcement learning with human feedback. *arXiv preprint arXiv:2502.17515*, 2025.

Heyang Zhao, Chenlu Ye, Quanquan Gu, and Tong Zhang. Sharp analysis for kl-regularized contextual bandits and rlhf. *arXiv preprint arXiv:2411.04625*, 2024.

Junhao Zheng, Shengjie Qiu, and Qianli Ma. Learn or recall? revisiting incremental learning with pre-trained language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14848–14877, 2024.

Qinqing Zheng, Shuxiao Chen, Qi Long, and Weijie Su. Federated f-differential privacy. In *International conference on artificial intelligence and statistics*, pages 2251–2259. PMLR, 2021.

Yanxin Zheng, Wensheng Gan, Zefeng Chen, Zhenlian Qi, Qian Liang, and Philip S Yu. Large language models for medicine: a survey. *International Journal of Machine Learning and Cybernetics*, 16(2):1015–1040, 2025.

Huiying Zhong, Zhun Deng, Weijie J Su, Zhiwei Steven Wu, and Linjun Zhang. Provable multi-party reinforcement learning with diverse human feedback. *arXiv preprint arXiv:2403.05006*, 2024.

Hongyi Zhou, Kai Ye, Erhan Xu, Jin Zhu, Shijin Gong, and Chengchun Shi. Demystifying group relative policy optimization: Its policy gradient is a u-statistic. *arXiv preprint arXiv:2603.01162*, 2026.

Xingyu Zhou, Yulian Wu, and Francesco Orabona. A unified theoretical analysis of private and robust offline alignment: from rlhf to dpo. *arXiv preprint arXiv:2505.15694*, 2025a.

Xingyu Zhou, Yulian Wu, Wenqian Weng, and Francesco Orabona. Square χpo: Differentially private and robust χ^2-preference optimization in offline direct alignment. *arXiv preprint arXiv:2505.21395*, 2025b.

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