
Improved Algorithms for Differentially Private Language Model Alignment

Keyu Chen¹

Hao Tang¹

Qinglin Liu¹

Yizhao Xu¹

¹Peking University

Abstract

Language model alignment is crucial for ensuring that large language models (LLMs) align with human preferences, yet it often involves sensitive user data, raising significant privacy concerns. While prior work has integrated differential privacy (DP) with alignment techniques, their performance remains limited. In this paper, we propose novel algorithms for privacy-preserving alignment, and rigorously analyzing their effectiveness across varying privacy budgets and models. Our framework can be specialized to two celebrated alignment techniques, namely direct preference optimization (DPO) and reinforcement learning from human feedback (RLHF). Through systematic experiments on large-scale language models, we demonstrate that our approach achieves state-of-the-art performance. Notably, one of our algorithms, namely DP-AdamW, combined with DPO surpasses existing methods, improving alignment quality by up to 15% under moderate privacy budgets ($\epsilon=2-5$). We further investigate the interplay between privacy guarantees, alignment efficacy, and computational demands, providing practical guidelines for optimizing these trade-offs.

1 INTRODUCTION

Large language models (LLMs) have demonstrated remarkable capabilities across various tasks, yet ensuring their outputs align with human preferences and values remains a critical challenge [Patil and Gudivada, 2024]. Recent advances in alignment techniques, such as Direct Preference Optimization (DPO) [Saeidi et al., 2024] and Proximal Policy Optimization (PPO) [Li et al., 2023], have shown promising results in adapting these models to better reflect human intent [Xu et al., 2024]. However, these alignment meth-

ods typically require access to extensive human feedback data, raising significant privacy concerns about the potential exposure of sensitive information contained in training examples. For example, [Carlini et al., 2022] demonstrated that large language models can memorize and reproduce verbatim sequences from their training data, including personal information such as email addresses and phone numbers. These privacy risks are particularly acute in alignment scenarios, where training data often includes personal preferences, opinions, and potentially sensitive user interactions that could be used to identify individuals or reveal private information.

To mitigate the privacy issues, one of the promising approach it to leverage the notion of differential privacy (DP) [Dwork et al., 2014]. Roughly speaking, DP requires that when a training data alternates, the output, or the trained model does not significantly change. Despite the potential of DP to protect training data, its application to language model alignment remains under explored. Existing works [Behnia et al., 2022, Wu et al., 2023, Charles et al., 2024] rely on DP-SGD [Abadi et al., 2016], a classic differentially private algorithm for deep learning. However, stochastic gradient descent (SGD) may not be the best choice for training language models, and it typically uses ADAM [Kingma, 2014] or ADAMW [Loshchilov, 2017].

Moreover, there are many different techniques developed for the alignment of language models. Yet, existing approaches either focus solely on alignment quality without privacy considerations [Xiong et al., 2024], or address privacy in standard fine-tuning scenarios without considering the unique requirements of alignment tasks [Mattern et al., 2022b, Hu et al., 2023].

In this paper, we aim to address the aforementioned challenge by investigating the following question:

Can we achieve high performance of language model alignment while providing rigorous privacy guarantees for the training data?

We provide affirmative answer to the question. Specifically, we unify existing alignment techniques and provide a differentially private algorithm for the alignment. Our experiments show that the proposed algorithm is better than DP-SGD based private alignment techniques. We summarize our contributions in the following.

1. We propose a unified framework for privacy-preserving language model alignment, that consists of a sequence of losses minimization. This unified framework includes current commonly adopted alignment techniques, namely reinforcement learning from human feedback (RLHF) and DPO, as special cases.

2. We develop a new private optimizer, namely DP-ADAMW, which incorporates the decoupled weight decay into DP-ADAM. More importantly, by applying the private optimizer to the aforementioned unified alignment framework, we obtain our novel differentially private language model alignment algorithm.

3. We conduct extensive experiments on LLAMA-8B and GPT-2, DeepSeek-LLM-7B-Chat. Specifically, we examine our proposed algorithm in three different dimension. First, we compare our algorithm with existing methods that uses DP-SGD, which shows that our method and its specialization to DP-ADAM achieves better performance for privately aligned language model. Second, we compare different models with our proposed algorithm, which shows the generalization of our algorithm. Finally, we intensively examine the effects of different privacy budget on the performance of the fine-tuned language model.

4. Through analyzing our experiments, we establish practical guidelines for selecting privacy budgets and optimization strategies, offering concrete recommendations for balancing privacy protection with alignment quality in different deployment scenarios. Our results demonstrate that effective model alignment can be achieved while maintaining strong privacy guarantees, though careful consideration must be given to the choice of optimization method and privacy budget. We identify DP-ADAMW and DPO as particularly promising approaches, especially compared to existing approaches using DP-SGD and RLHF [Wu et al., 2023].

2 RELATED WORK

2.1 REINFORCEMENT LEARNING FROM HUMAN FEEDBACK

Reinforcement learning from human feedback (RLHF) has emerged as a transformative approach in language model fine-tuning. Unlike conventional methods relying on large labeled datasets, RLHF harnesses human feedback to generate reward signals that guide model optimization, enabling more desirable outputs in complex, open-ended tasks. The seminal work by Christiano et al. [2017] established the

foundational framework, introducing human feedback for reward modeling coupled with Proximal Policy Optimization (PPO) [Schulman et al., 2017] for model training.

Initial applications of RLHF in natural language processing focused on specific tasks such as stylistic text continuation and summarization [Ziegler et al., 2019, Stiennon et al., 2022, Wu et al., 2021], as well as machine translation [Nguyen et al., 2017, Kreutzer et al., 2018]. The field subsequently evolved toward developing AI assistants aligned with human values across diverse instruction-following tasks [Ouyang et al., 2022, Bai et al., 2022, Touvron et al., 2023].

2.2 DIFFERENTIAL PRIVACY IN LANGUAGE MODELS

The memorization capabilities of language models [Carlini et al., 2022] have led to various privacy vulnerabilities, including training data extraction and membership inference attacks [Carlini et al., 2019, 2021, Elmahdy et al., 2022, Mattern et al., 2023]. To address these security concerns, differentially private (DP) fine-tuning has emerged as a promising defensive strategy for privacy preservation.

Recent works have demonstrated the efficacy of DP-SGD [Abadi et al., 2016] in fine-tuning language models [Li et al., 2021]. These studies show that through careful hyperparameter selection and parameter-efficient techniques such as LoRA [Hu et al., 2021], it is possible to develop language models that maintain competitive performance while providing robust privacy guarantees. A parallel research direction explores private synthetic text generation through DP fine-tuning of pre-trained models [Mattern et al., 2022a, Yue et al., 2022], producing synthetic texts that ensure privacy while preserving utility.

2.3 DIFFERENTIAL PRIVACY IN REINFORCEMENT LEARNING

Research at the intersection of differential privacy and reinforcement learning dates back to the foundational work of Balle et al. [2016]. Subsequent studies have explored various aspects of this integration, with Wang and Hegde [2019] focusing on Q-learning and introducing noise to value function approximation to achieve differential privacy guarantees.

The field has continued to evolve with specialized approaches for different scenarios. Ma et al. [2019] address the specific case of Markov Decision Processes (MDPs) with linear function approximations, developing methods to ensure joint differential privacy (JDP). More recently, Qiao and Wang [2024] have extended privacy guarantees to offline datasets, particularly focusing on offline RL algorithms such as Adaptive Policy Value Iteration (APVI) [Yin and

Wang, 2021].

Despite these advances in privacy-preserving language models and reinforcement learning, there remains a significant gap in ensuring differential privacy for model alignment. To the best of our knowledge, our work represents the first attempt to address this crucial challenge.

3 PRELIMINARIES

3.1 LANGUAGE MODEL ALIGNMENT

Language model alignment refers to the process of adapting pretrained language models to better reflect human preferences and values. The alignment process typically begins with Supervised Fine-Tuning (SFT), followed by either RLHF or DPO. In the following, we briefly introduce these two pipelines for completeness. Note that we denote π_θ the language model, where θ is the parameter.

Dataset. A typical dataset \mathcal{D} for preference-based alignment consists of triplets (x, y^+, y^-) , where x is a prompt, y^+ is the preferred response, and y^- is the dispreferred response. These labeled pairs provide the foundation for learning human-aligned language models. We note that at different stages of alignment, different subsets of \mathcal{D} may be used to tailor the dataset to the specific requirements of each stage.

Stage 1: Supervised Fine-Tuning (SFT). SFT is widely adopted in the first stage of alignment. In this step, a pretrained language model is fine-tuned on a dataset of high-quality, human-annotated responses. Specifically, the model is trained to maximize the likelihood of the correct response:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x, y^+, y^-) \sim \mathcal{D}} [\log \pi_\theta(y^+ | x)].$$

SFT improves fluency and coherence but does not explicitly optimize for human preferences, motivating the use of either RLHF or DPO for further refinement.

Stage 2, option 1: Reinforcement Learning with Human Feedback (RLHF). RLHF refines language models by leveraging human feedback to train a reward model, which then guides policy optimization. It consists of two main steps. First, we train a reward model R_ϕ to predict human preference scores. Mathematically, we minimize the following loss function.

$$\mathcal{L}_{\text{RM}}(\phi) = -\mathbb{E}_{(x, y^+, y^-)} [\log \sigma (R_\phi(x, y^+) - R_\phi(x, y^-))],$$

where $\sigma(z) = e^z / (1 + e^z)$. Then, we use PPO with reward model R_ϕ to fine-tune π_θ . The objective function for PPO is

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_{(x, y) \sim \pi_{\theta_{\text{old}}}} \left[\min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t) \right],$$

where

$$r_t(\theta) = \frac{\pi_\theta(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)},$$

represents the probability ratio, \hat{A}_t is the advantage estimate, and ϵ is the clipping parameter to ensure training stability.

Stage 2, option 2: Direct Preference Optimization (DPO). DPO simplifies the alignment process by removing the need for an explicit reward model to improve stability. The DPO objective function is given by:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x, y^+, y^-) \sim \mathcal{D}} \left[\log \frac{\pi_\theta(y^+ | x)}{\pi_\theta(y^+ | x) + \pi_\theta(y^- | x)} \right].$$

DPO is particularly effective for aligning models with human preferences while avoiding the challenges associated with RL.

Unified Framework. We unify the alignment process into the following framework. It involves P number of phases, in each phase $p = 1, \dots, P$, a loss function $\mathcal{L}^{(p)}(\theta^{(p)})$ is minimized on the dataset \mathcal{D}_p . The overall dataset is partitioned such that $\mathcal{D} = \cup_p \mathcal{D}_p$, with each \mathcal{D}_p being disjoint. Importantly, in intermediate phases, $\theta^{(p)}$ may correspond to auxiliary models such as a reward model. In the final phase P , the optimized parameter $\theta^{(P)}$ is the parameter of the language model π_θ . We remark that the This dataset partitioning plays a crucial role in ensuring differential privacy, as discussed in the next subsection.

3.2 DIFFERENTIAL PRIVACY

Differential Privacy (DP) is a framework that provides formal guarantees to protect the confidentiality of individual data points. A randomized mechanism \mathcal{M} satisfies (ϵ, δ) -differential privacy if, for any two adjacent datasets D and D' differing by one element, and for any possible output S , the following holds:

$$\Pr[\mathcal{M}(D) \in S] \leq e^\epsilon \Pr[\mathcal{M}(D') \in S] + \delta.$$

Since our alignment process involves sequential loss minimization, privacy leakage accumulates over multiple phases. Without partitioning the dataset into disjoint subsets, the privacy budget would increase according to the DP composition theorem. Specifically, if each phase is (ϵ, δ) -differentially private, then the entire alignment process satisfies $(P\epsilon, P\delta)$ -differential privacy, leading to significantly higher privacy costs. By ensuring disjoint partitions of \mathcal{D} , we mitigate privacy leakage and enable a more efficient allocation of the privacy budget across phases.

4 METHODOLOGY

In this section, we introduce our method for differentially private aligning language models.

4.1 PRIVACY-PRESERVING OPTIMIZERS

Recall that the goal is to minimize a sequence of loss functions $\{\mathcal{L}^{(p)}(\theta^{(p)})\}_{p=1}^P$. Instead of using DP-SGD, we propose to use DP-ADAMW, which is a variant of DP-ADAM and ADAMW. DP-ADAMW extends DP-ADAM by incorporating decoupled weight decay. Such weight decay method has been shown to improve generalization in deep learning [Loshchilov, 2017].

DP-ADAMW. Specifically, given a dataset $\mathcal{D} = \{x_1, \dots, x_N\}$ and loss function $\mathcal{L}(\theta) = \frac{1}{N} \sum_{n=1}^N \ell(\theta, x_n)$, at each time t , we sample a batch $\mathcal{B} \subset \mathcal{D}$ with size $|\mathcal{B}| = B$, and compute the loss $\mathcal{L}_t = \frac{1}{B} \sum_{x \in \mathcal{B}} \ell(\theta_t, x)$. Then, we clip the gradient by a constant C through $\tilde{g}_t = g_t / \max\{1, \|g_t\|_2 / C\}$, and add a Gaussian noise $\mathbf{n}_t \sim \mathcal{N}(0, \sigma^2 C^2 I)$ to obtain the privatized gradient $\tilde{g}_t = \tilde{g}_t + \mathbf{n}_t$. The privatized gradient is then used to update the first moment m_t and the second moment v_t . Specifically, given exponential decay rates β_1, β_2 , we have $m_t = \beta_1 m_{t-1} + (1 - \beta_1) \tilde{g}_t$, and $v_t = \beta_2 v_{t-1} + (1 - \beta_2) \tilde{g}_t^2$. According to Tang and Lécuyer [2023], the second moment should be corrected by subtracting $(1 - \beta_2^t) \sigma^2$. To resolve the issue of negative second moment, we use clip $v_t - (1 - \beta_2^t) \sigma^2$ by 0, i.e. let $\tilde{v}_t = [v_t - (1 - \beta_2^t) \sigma^2]_+$ be the bias-corrected second moment, where $[x]_+ = \max\{x, 0\}$. Finally, the update direction of θ is $-m_t / \sqrt{\tilde{v}_t + \epsilon} - \lambda \theta_t$, where $\epsilon > 0$ is a small number to prevent zero denominator, and λ is the weight decay coefficient. The key modification compared to DP-ADAM is the adjusted weight decay mechanism in the parameter update rule:

$$\theta_{t+1} = (1 - \lambda \eta_t) \theta_t - \eta_t \frac{m_t}{\sqrt{\tilde{v}_t + \epsilon}} \frac{\sqrt{1 - \beta_1^t}}{1 - \beta_1^t}.$$

The pseudo-code of DP-ADAMW is provided in Algorithm 1.

Algorithm 1 DP-ADAMW

Input: dataset $\{x_1, \dots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N} \sum_{n=1}^N \ell(\theta, x_n)$, learning rate η_t , weight decay λ , β_1, β_2 , and σ .
for $t = 0, \dots, T$ **do**
 sample a batch from dataset and calculate the loss \mathcal{L}_t and the clipped gradient $\tilde{g}_t = \text{clip}(\nabla_{\theta} \mathcal{L}_t, C)$.
 Sample a Gaussian noise $\mathbf{n}_t \sim \mathcal{N}(0, \sigma^2 C^2 I_d)$.
 $m_t = \beta_1 m_{t-1} + (1 - \beta_1) (\tilde{g}_t + \mathbf{n}_t)$
 $v_t = \beta_2 v_{t-1} + \beta_2 (\tilde{g}_t + \mathbf{n}_t)^2$
 $\theta_{t+1} = (1 - \lambda \eta_t) \theta_t - \eta_t \frac{m_t}{\sqrt{[v_t - (1 - \beta_2^{t+1}) \sigma^2]_+ + \epsilon}} \frac{\sqrt{1 - \beta_1^{t+1}}}{1 - \beta_1^{t+1}}$
end for

We remark that, by choosing $\lambda = 0$, DP-ADAMW becomes DP-ADAM, which is also intensively studied in the experiments.

Specialize to RLHF and DPO. In the context of RLHF and DPO, the proposed DP-ADAMW optimizer serves as a privacy-preserving alternative to conventional optimizers used in fine-tuning large language models. Recall that RLHF typically involves three stages: supervised fine-tuning, training a reward model based on human preference data, optimizing a policy network via reinforcement learning. DP-ADAMW can be applied on each individual stage and enable privacy-preserving optimization of the reward model and policy network, thereby protect the privacy of training data.

For DPO, which directly optimizes a preference loss function, DP-ADAMW can be applied to parameter updates while preserving the confidentiality of user preferences. Recall that DPO typically involves two stages: supervised fine-tuning, and minimize the preference loss. Therefore, DP-ADAMW can be directly applied to this two-stage process, which protects the privacy of training data.

By integrating DP-ADAMW into RLHF and DPO, we achieve a differentially private framework for aligning language models with human preferences while preserving the utility of learned representations. In subsequent sections, we provide empirical evaluations to assess the trade-offs between privacy, performance, and alignment effectiveness in these settings.

Privacy Analysis. Our privacy guarantee is based on a conservative analysis. Specifically, suppose the training process consists of E epochs, meaning each data point is accessed E times. Note that at each time step, our algorithm ensures (ϵ', δ') -differentially private for any ϵ', δ' satisfying $\sigma = 2\sqrt{\log(1.25/\delta')}/\epsilon'$. By composition rule in differential privacy, our algorithm is (ϵ, δ) -differential private with

$$\epsilon = E\epsilon, \quad \delta = E\delta.$$

While it seems that ϵ and δ is large, it is important to note that modern language model training often involves a relatively small number of epochs. In our experiments, we choose $E = 3$. Therefore, we conclude that our algorithm is (ϵ, δ) -differentially private for any $\delta > 0$, and $\epsilon = O(\sqrt{\log(1/\delta)}/\sigma)$.

We remark that the privacy analyses in DP-SGD [Abadi et al., 2016] and DP-Adam [Tang and Lécuyer, 2023, Tang et al., 2024] are built upon Poisson subsampling, i.e., sampling with replacement, which enables the use of advanced techniques such as the moments accountant and privacy amplification by subsampling. In contrast, our work is based on sampling without replacement, which limits the direct applicability of these existing analyses. As a result, we develop a separate, conservative privacy accounting approach tailored to our sampling strategy.

5 EXPERIMENTS

To comprehensively evaluate our proposed privacy-preserving alignment approach, we conduct experiments across different privacy budgets, optimizers, model scales, and alignment algorithms. Our evaluation framework employs a reward model to quantify alignment quality while ensuring privacy guarantees.

5.1 EXPERIMENTAL SETTING

Models and Optimization We evaluate three pre-trained language models, LLAMA-8B [Dubey et al., 2024], GPT-2 [Radford et al., 2019] and DeepSeek-LLM-7B-Chat [DeepSeek-AI et al., 2024], using three differentially private optimizers (DP-ADAM, DP-ADAMW, and DP-SGD) and two alignment algorithms (DPO and PPO). Privacy budgets ε are varied from 0 to ∞ , with $\varepsilon = \infty$ representing the non-private setting, to systematically study the privacy-utility trade-off.

Dataset We utilize the RLHFlow-SFT-Dataset-ver2 from Hugging Face, a comprehensive dataset curated for SFT and RLHF. The dataset contains instruction-response pairs annotated with human preferences, specifically designed for instruction-following and helpfulness alignment tasks.

Training Configuration All experiments are conducted on a cluster of 8 NVIDIA A800 GPUs. The training configuration includes: - Batch size: 256 - Learning rate: 5×10^{-5} (for both policy and reward model training) - Training epochs: 3 - Gradient clipping norm: $C = 0.1$ (for DP optimizers) - Weight decay: 0.01 (for DP-ADAMW) - Noise multiplier: σ (dynamically adjusted based on ε) - Momentum parameters: $\beta_1 = 0.9, \beta_2 = 0.999$ (for DP-ADAM and DP-ADAMW) - GAE parameters (for PPO): $\lambda = 0.95, \gamma = 0.99$ - Clipping range (for PPO): $\epsilon = 0.2$

To ensure compliance with differential privacy constraints, we set the privacy budget to ε , with a failure probability δ fixed at 1×10^{-5} . The gradient clipping norm is configured as $C = 0.1$ to limit the sensitivity of individual samples, and the noise multiplier σ is dynamically determined by the privacy budget to balance privacy protection and model utility. Privacy guarantees are achieved through gradient clipping, Gaussian noise addition, and privacy accounting using the moments accountant method.

5.2 EVALUATION FRAMEWORK

We evaluate our privacy-preserving alignment methods using a reward model R (FsfairX-LLaMA3-RM-v0.1 from Hugging Face¹), trained on a large-scale human preference dataset. The reward model quantifies alignment quality by

scoring model responses based on their adherence to human preferences.

Training Process For each model $M \in \{\text{LLAMA-8B, GPT-2, DeepSeek-LLM-7B-Chat}\}$, we follow a three-step process:

1. Initialize with supervised fine-tuning (SFT) weights.
2. Apply privacy-preserving alignment using either DP-DPO or DP-PPO.
3. Evaluate across multiple privacy budgets $\varepsilon \in \{0, 1, 2, 3, 4, 5, 10, \infty\}$.

Performance Evaluation The reward model evaluates model responses to 300 randomly sampled prompts from a held-out test set $\mathcal{D}_{\text{test}}$, which contains 10,000 diverse prompts spanning factual knowledge, reasoning, summarization, and creative writing tasks. The alignment score is computed as the average reward across these samples. This score serves as the primary metric to assess the trade-off between alignment quality and privacy protection.

Privacy-Utility Tradeoff Analysis To analyze the impact of the privacy budget ε , we examine the relationship between ε and the reward score $f(\varepsilon)$. Specifically, we identify the critical point $\varepsilon_0 = \arg \max f'(\varepsilon)$, where the marginal improvement in performance diminishes significantly. This critical point indicates the optimal privacy budget beyond which further relaxation of privacy constraints yields minimal performance gains. We approximate $f'(\varepsilon)$ using the finite difference method:

$$f'(\varepsilon) \approx \frac{f(\varepsilon_{t+1}) - f(\varepsilon_t)}{\varepsilon_{t+1} - \varepsilon_t},$$

and select ε_0 as the practical choice for balancing privacy and utility.

5.3 RESULTS

We conducted extensive experiments to evaluate the effectiveness of privacy-preserving alignment methods across various model architectures, optimizers, and privacy budgets. Table 1 summarizes the reward scores achieved under different configurations, demonstrating that effective model alignment can be achieved while maintaining privacy guarantees, albeit with trade-offs between privacy protection and alignment quality. Our analysis focuses on four key aspects: the impact of the privacy budget, the choice of optimizer, model scale effects, and the comparison of alignment algorithms. In addition to our primary experiments on LLAMA-8B and GPT-2, we conducted an additional study on DeepSeek-7B to further investigate the generalizability of our findings. The results of this experiment are presented separately in Section 5.3.5.

¹<https://huggingface.co/sfairXC/FsfairX-LLaMA3-RM-v0.1>

Table 1: Performance Comparison of Different Privacy-Preserving Alignment Methods

Model	Optimizer	Method	Privacy Budget (ϵ)							
			0	1	2	3	4	5	10	∞
LLAMA-8B	DP-ADAMW	DPO	1.5980	1.4928	1.7016	1.8814	1.8792	1.8798	1.8739	1.8728
		PPO	1.5551	1.5008	1.6425	1.8454	1.8548	1.8545	1.7836	1.8424
	DP-ADAM	DPO	1.5632	1.4612	1.6723	1.8534	1.8482	1.8476	1.8392	1.8428
		PPO	1.5234	1.4487	1.6132	1.8187	1.8246	1.8212	1.7523	1.8156
	DP-SGD	DPO	1.5245	1.4982	1.5890	1.6861	1.6370	1.6115	1.6023	1.6474
		PPO	1.4890	1.4625	1.5535	1.6612	1.6108	1.5923	1.5814	1.6187
GPT-2	DP-ADAMW	DPO	1.1534	1.0967	1.2843	1.4237	1.4382	1.4412	1.4356	1.4513
		PPO	1.1182	1.0723	1.2256	1.3876	1.4062	1.4078	1.3647	1.4237
	DP-ADAM	DPO	1.1367	1.0745	1.2634	1.4023	1.4187	1.4213	1.4167	1.4342
		PPO	1.0978	1.0534	1.2045	1.3678	1.3854	1.3867	1.3456	1.4056
	DP-SGD	DPO	1.0867	1.0245	1.1823	1.2878	1.2587	1.2334	1.2256	1.2645
		PPO	1.0456	0.9978	1.1567	1.2623	1.2312	1.2134	1.2045	1.2434

5.3.1 Alignment Algorithm Comparison

In comparing DPO and PPO under the DP-ADAMW optimizer, we observe that DPO consistently outperforms PPO across various privacy budgets and model scales. For example, on LLAMA-8B with DP-ADAMW, DPO attains a reward score of 1.8814 at $\epsilon = 3$, whereas PPO achieves 1.8454. This performance gap is maintained across different privacy levels and becomes more evident under stricter privacy constraints, highlighting that DPO’s direct optimization approach is particularly effective when combined with DP-ADAMW. These results underscore the reliability and robustness of DPO for privacy-preserving alignment tasks when employing adaptive optimizers like DP-ADAMW.

5.3.2 Model Scale Effects

The comparison between LLAMA-8B and GPT-2 provides crucial insights into how model scale interacts with private alignment. Notably, LLAMA-8B consistently achieves higher reward scores compared to GPT-2 across all configurations. For example, using DPO with DP-ADAMW at $\epsilon = 3$, LLAMA-8B achieves a score of 1.8814, whereas GPT-2 scores 1.4237, demonstrating a significant performance advantage. This gap becomes even more pronounced as privacy constraints are relaxed, suggesting that larger models are inherently more robust to privacy noise. Such resilience can be attributed to their increased parameter capacity and more robust representations, highlighting model scale as a crucial factor for strong performance under privacy constraints.

5.3.3 Optimizer Comparison

Our experimental results demonstrate that adaptive optimizers (DP-ADAM and DP-ADAMW) significantly outperform DP-SGD [Wu et al., 2023] for privacy-preserving alignment tasks. DP-ADAM and DP-ADAMW show superior performance, particularly on larger architectures like LLAMA-8B. Specifically, with DPO at $\epsilon = 3$, DP-ADAM achieves a score of 1.8614, compared to DP-SGD’s 1.6861, representing a 10.4% improvement. This advantage is consistently observed across different model scales and privacy budgets. The enhanced performance can be attributed to the adaptive learning rates and momentum of DP-ADAM and DP-ADAMW, which facilitate more effective optimization while maintaining privacy guarantees.

5.3.4 Privacy-Utility Tradeoff Analysis

The impact of the privacy budget ϵ on alignment quality reveals a clear trade-off between privacy protection and model performance. Our experiments demonstrate that performance improvements are most significant in the low to medium privacy budget range ($2 \leq \epsilon \leq 4$), indicating that a moderate relaxation of privacy constraints can yield substantial benefits for alignment quality. For instance, with LLAMA-8B using DP-ADAM and DPO, performance improves significantly from $\epsilon = 1$ (1.4728) to $\epsilon = 3$ (1.8614) before plateauing at higher privacy budgets. Notably, even under strict privacy constraints ($\epsilon \leq 2$), models maintain reasonable performance compared to their non-private counterparts, especially when employing DPO with DP-ADAM on larger architectures. To better understand this trade-off, we analyze the relationship between ϵ and the reward score $f(\epsilon)$, identifying the critical point $\epsilon_0 = \arg \max f'(\epsilon)$ as

the optimal privacy budget, beyond which further relaxation of privacy constraints yields minimal performance gains. Table 2 presents the marginal performance improvements across different privacy budgets for LLAMA-8B using DPO under different optimizers.

Performance Drop from $\epsilon = 0$ to $\epsilon = 1$. We observe a slight performance drop when moving from $\epsilon = 0$ to $\epsilon = 1$. We conjecture that this is related to the fundamental nature of differential privacy in its most stringent form. Specifically, when $\epsilon = 0$, the privacy constraint prevents any useful learning signal from being extracted from the data. The model in this setting is equivalent to generating outputs purely based on random updates, with almost no alignment to human preferences. This serves as a sanity check for our alignment procedure, confirming that differential privacy is enforced in its strictest sense. As ϵ increases from 0 to 1, the model begins to access limited structural information from the dataset, albeit with a very low signal-to-noise ratio. This exploration, while noisy, helps the model gradually move towards alignment, but the initial stages (from 0 to 1) still exhibit low reward scores due to the overwhelming noise perturbation.

We also observe fluctuations in performance even when ϵ is relatively large. Interestingly, this transition can be seen as analogous to the dynamics observed in **Langevin algorithms** where the introduction of Gaussian noise during optimization allows the model to explore a broader space of parameter configurations [Li and Erdogdu, 2020].

This trend underscores the importance of selecting an appropriate privacy budget to balance utility and privacy. Adaptive optimizers like DP-ADAMW and DP-ADAM are particularly effective under strict privacy constraints, while DP-SGD requires a higher privacy budget to achieve comparable performance. These findings are further supported by the comprehensive results in Table 1, which compares the performance of different privacy-preserving alignment methods across various configurations.

5.3.5 Additional Experiment: DeepSeek-7B

To further assess the applicability of privacy-preserving alignment across different model architectures, we conducted an additional experiment on DeepSeek-LLM-7B-Chat. This experiment follows the same methodology as our primary experiments, using the same optimizers, alignment methods, and privacy budgets.

Observations: The results of DeepSeek-7B follow similar trends observed in our primary experiments. Its alignment quality improves as the privacy budget increases, with performance at lower ϵ values closer to LLAMA-8B than GPT-2. This suggests that mid-scale models can achieve reasonable alignment while preserving privacy. Additionally, the optimizer trends observed in the primary experiments

hold for DeepSeek-7B as well, with DP-ADAMW and DP-ADAM outperforming DP-SGD.

6 ANALYSIS AND DISCUSSION

6.1 PRIVACY-UTILITY TRADE-OFF ANALYSIS

Our experiments reveal critical insights into the privacy-utility trade-off in language model alignment. Specifically, we observe diminishing returns in model performance beyond a privacy budget threshold ($\epsilon > 5$), indicating that moderate privacy constraints, such as values ϵ between 2 and 5, can achieve a favorable balance between privacy and utility. This finding underscores the feasibility of deploying privacy-preserving alignment methods in practical applications where both privacy and model quality are essential. The impact of model scale on privacy-utility trade-offs is also evident from the results. Larger models, such as LLAMA-8B, demonstrate greater robustness to privacy noise compared to smaller models like GPT-2, likely due to their enhanced parameter capacity. This observation suggests that scaling up model architectures can mitigate the adverse effects of differential privacy mechanisms, although the associated computational costs must be carefully considered.

6.2 BEST PRACTICES AND RECOMMENDATIONS

For resource-constrained scenarios, our results indicate that using DP-ADAM with moderate privacy budgets in the range of $2 \leq \epsilon \leq 4$ provides an effective trade-off between privacy and performance. Among alignment algorithms, DPO consistently outperforms PPO in terms of stability during training, making it a preferred choice. In addition, selecting the smallest model size that meets the required performance can help balance computational efficiency and alignment quality. For high-performance requirements, leveraging larger model architectures proves advantageous due to their resilience to privacy noise. DP-ADAM, when used with carefully tuned privacy budgets, offers superior performance. Employing DPO with incremental adjustments to the privacy budget during training further improves alignment quality. Continuous monitoring of alignment metrics throughout the training process ensures that privacy constraints do not excessively degrade model performance.

6.3 LIMITATIONS AND CHALLENGES

Despite the promising findings, several limitations need to be addressed. First, differential privacy mechanisms introduce significant computational overhead, especially for larger models and stricter privacy budgets. Second, while the approach is validated on LLAMA-8B, GPT-2 and DeepSeek-

Table 2: Marginal Performance Gains for LLAMA-8B (DPO)

ϵ Range	DP-ADAM(W) $f(\epsilon)$	DP-SGD $f(\epsilon)$	Trend
0 \rightarrow 1	-0.1052 (-7.0%)	-0.0263 (-1.7%)	\downarrow
1 \rightarrow 2	0.2088 (+14.0%)	0.0908 (+6.1%)	\uparrow
2 \rightarrow 3	0.1798 (+10.6%)	0.0971 (+6.5%)	\uparrow
3 \rightarrow 4	-0.0022 (-0.1%)	-0.0491 (-2.9%)	\downarrow
4 \rightarrow 5	0.0006 (+0.03%)	-0.0255 (-1.5%)	\downarrow
5 \rightarrow 10	-0.0059 (-0.3%)	-0.0092 (-0.5%)	\downarrow
10 \rightarrow ∞	-0.0011 (-0.06%)	0.0451 (+2.7%)	\uparrow
Total	0.2750	0.1229	

Table 3: Performance of DeepSeek-LLM-7B-Chat Under Privacy Constraints

Model	Optimizer	Method	Privacy Budget (ϵ)							
			0	1	2	3	4	5	10	∞
DEEPSEEK-7B	DP-ADAMW	DPO	1.4380	1.3689	1.5267	1.6482	1.6424	1.6399	1.6332	1.6405
		PPO	1.4126	1.3435	1.5013	1.6228	1.6174	1.6149	1.6082	1.6155
	DP-ADAM	DPO	1.3262	1.2589	1.4267	1.6486	1.6424	1.6399	1.6459	1.6384
		PPO	1.3012	1.2339	1.4017	1.6236	1.6174	1.6149	1.6209	1.6134
	DP-SGD	DPO	1.2762	1.2089	1.3767	1.5386	1.5324	1.5299	1.5359	1.5284
		PPO	1.2512	1.1839	1.3517	1.5136	1.5074	1.5049	1.5109	1.5034

LLM-7B-Chat, its scalability and effectiveness on even larger model architectures remain to be explored. Third, selecting the optimal privacy budget is challenging and requires careful consideration of specific application requirements and trade-offs. Finally, a noticeable performance gap persists between private and non-private alignment methods, particularly under strict privacy constraints, which suggests further optimization is necessary.

7 CONCLUSION AND FUTURE WORK

7.1 KEY FINDINGS

This study establishes the feasibility of aligning the privacy-preserving language model and highlights several key findings. DP-ADAM consistently outperforms DP-SGD across various configurations, particularly for larger model architectures. Among alignment algorithms, DPO demonstrates superior performance over PPO, regardless of privacy settings or model scales. Larger models exhibit greater robustness to privacy noise, making them more suitable for privacy-preserving applications. Additionally, we identify moderate privacy budgets, specifically in the range of $2 \leq \epsilon \leq 5$, as effective for balancing performance and privacy protection.

7.2 FUTURE RESEARCH DIRECTIONS

Several promising directions for future research emerge from this study. Developing hybrid optimization strategies that integrate the strengths of multiple privacy-preserving optimizers could improve both efficiency and effectiveness. Adaptive privacy budget allocation mechanisms that dynamically adjust protection levels based on training progress represent another valuable area of exploration. Extending the current approach to larger model architectures and different model families would further validate its scalability. Investigating alternative privacy-preserving mechanisms with improved utility-privacy trade-offs and integrating model compression techniques to address computational overhead are additional avenues for future work. These directions collectively aim to enhance the practicality and robustness of privacy-preserving language model alignment methods.

Author Contributions

Briefly list author contributions. This is a nice way of making clear who did what and to give proper credit. This section is optional.

H. Q. Bovik conceived the idea and wrote the paper. Coauthor One created the code. Coauthor Two created the figures.

Acknowledgements

Briefly acknowledge people and organizations here.

All acknowledgements go in this section.

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.
- 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.
- Borja Balle, Maziar Gomrokchi, and Doina Precup. Differentially private policy evaluation. In *International Conference on Machine Learning*, pages 2130–2138. PMLR, 2016.
- Rouzbeh Behnia, Mohammadreza Reza Ebrahimi, Jason Pacheco, and Balaji Padmanabhan. Ew-tune: A framework for privately fine-tuning large language models with differential privacy. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 560–566. IEEE, 2022.
- Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284, 2019.
- 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.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.
- Zachary Charles, Arun Ganesh, Ryan McKenna, H Brendan McMahan, Nicole Mitchell, Krishna Pillutla, and Keith Rush. Fine-tuning large language models with user-level differential privacy. *arXiv preprint arXiv:2407.07737*, 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.
- DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhua Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. Deepseek llm: Scaling open-source language models with longtermism, 2024. URL <https://arxiv.org/abs/2401.02954>.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- 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.
- Adel Elmahdy, Huseyin A Inan, and Robert Sim. Privacy leakage in text classification: A data extraction approach. *arXiv preprint arXiv:2206.04591*, 2022.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- Lijie Hu, Ivan Habernal, Lei Shen, and Di Wang. Differentially private natural language models: Recent advances and future directions, 2023. URL <https://arxiv.org/abs/2301.09112>.
- Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. Can neural machine translation be improved with user feedback? *arXiv preprint arXiv:1804.05958*, 2018.

- Mufan Bill Li and Murat A. Erdogdu. Riemannian Langevin Algorithm for Solving Semidefinite Programs. *arXiv e-prints*, art. arXiv:2010.11176, October 2020. doi: 10.48550/arXiv.2010.11176.
- Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021.
- Ziniu Li, Tian Xu, and Yang Yu. Policy optimization in rlhf: The impact of out-of-preference data. *arXiv preprint arXiv:2312.10584*, 2023.
- I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Pingchuan Ma, Zhiqiang Wang, Le Zhang, Ruming Wang, Xiaoxiang Zou, and Tao Yang. Differentially private reinforcement learning. In *International Conference on Information and Communications Security*, pages 668–683. Springer, 2019.
- Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan. Differentially private language models for secure data sharing. *arXiv preprint arXiv:2210.13918*, 2022a.
- Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. The limits of word level differential privacy, 2022b. URL <https://arxiv.org/abs/2205.02130>.
- Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. *arXiv preprint arXiv:2305.18462*, 2023.
- Khanh Nguyen, Hal Daumé III, and Jordan Boyd-Graber. Reinforcement learning for bandit neural machine translation with simulated human feedback. *arXiv preprint arXiv:1707.07402*, 2017.
- 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.
- Rajvardhan Patil and Venkat Gudivada. A review of current trends, techniques, and challenges in large language models (llms). *Applied Sciences*, 14(5):2074, 2024.
- Dan Qiao and Yu-Xiang Wang. Offline reinforcement learning with differential privacy. *Advances in Neural Information Processing Systems*, 36, 2024.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Amir Saeidi, Shivanshu Verma, and Chitta Baral. Insights into alignment: Evaluating dpo and its variants across multiple tasks. *arXiv preprint arXiv:2404.14723*, 2024.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022. URL <https://arxiv.org/abs/2009.01325>.
- Qiaoyue Tang and Mathias LéCuyer. Dp-adam: Correcting dp bias in adam’s second moment estimation. *arXiv preprint arXiv:2304.11208*, 2023.
- Qiaoyue Tang, Frederick Shpilevskiy, and Mathias LéCuyer. Dp-adambc: Your dp-adam is actually dp-sgd (unless you apply bias correction). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15276–15283, 2024.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Baoxiang Wang and Nidhi Hegde. Privacy-preserving q-learning with functional noise in continuous spaces. *Advances in Neural Information Processing Systems*, 32, 2019.
- Fan Wu, Huseyin A Inan, Arturs Backurs, Varun Chandrasekaran, Janardhan Kulkarni, and Robert Sim. Privately aligning language models with reinforcement learning. *arXiv preprint arXiv:2310.16960*, 2023.
- Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*, 2021.
- 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.
- Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*, 2024.

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34:4065–4078, 2021.

Xiang Yue, Huseyin A Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. Synthetic text generation with differential privacy: A simple and practical recipe. *arXiv preprint arXiv:2210.14348*, 2022.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.