

PRIVATE DATA SYNTHESIS FOR PREFERENCE ALIGNMENT OF LARGE LANGUAGE MODELS

Anonymous authors

Paper under double-blind review

ABSTRACT

Preference alignment has become a crucial technique for aligning large language models (LLMs) with human values. However, training on real human preference data raises privacy concerns, as these datasets often contain sensitive user prompts and human judgments. To address this, we propose **DPPrefSyn**, a novel algorithm for generating differentially private (DP) synthetic preference data to enable privacy-preserving preference alignment. DPPrefSyn addresses three key challenges: modeling diverse human preferences via DP clustering and per-cluster DP scoring models; reducing dimensionality with DP-PCA to improve efficiency; and conserving privacy budget by leveraging public prompts. We conduct extensive experiments on three standard benchmarks and compare our method with DP fine-tuning on real data. Our results show that our framework achieves competitive performance under strong privacy guarantees. These results open up new possibilities for preference alignment with privacy protection for a broad range of applications. To the best of our knowledge, this is the first work to generate DP synthetic preference data for LLM alignment.

1 INTRODUCTION

Preference alignment algorithms, such as RLHF (Stiennon et al., 2020) and DPO (Rafailov et al., 2024), are widely used to align large language models (LLMs) with human expectations. This is typically achieved by collecting a dataset of pairwise preferences, where human annotators indicate which of two responses better answers a given prompt. Such preference data is then used to fine-tune LLMs by encouraging the model to rank preferred responses over less preferred ones (Ziegler et al., 2019; Ouyang et al., 2022; Bai et al., 2022). The effectiveness and reliability of preference alignment have motivated its adoption in applications such as chat assistants (Achiam et al., 2023), mathematical reasoning tools (Shao et al., 2024), and code generators (Shen et al., 2023).

However, privacy is a significant concern when aligning LLMs with human preference data, as it contains real user prompts and feedback. These prompts may disclose personal information related to health, identity, or other sensitive topics, and human feedback can reveal private beliefs, preferences, or behavioral patterns (Li et al., 2023; Yu et al., 2024). To address this concern, existing works (Chowdhury et al., 2024; Yu et al., 2024; Wu et al., 2023a) have explored approaches to mitigate these risks using the rigorous privacy safeguards provided by differential privacy (DP) (Dwork et al., 2006b) and shown encouraging results. However, many of these approaches protect only part of the data, for example, by privatizing either the user prompts (Yu et al., 2024) or the preference labels (Chowdhury et al., 2024; Zhang et al., 2025), but not both. Such partial protection leaves potential privacy gaps, as adversaries could still infer sensitive attributes from unprotected prompts or preference annotations. One line of work focuses on specific post-training algorithms, such as PPO (Wu et al., 2023a), and is not compatible with newer methods like DPO. Moreover, these methods (Yu et al., 2024; Chowdhury et al., 2024; Zhang et al., 2025; Wu et al., 2023a) are typically constrained by limited private preference datasets, and often fall short in achieving high-quality preference alignment, due to the high cost and scalability challenges associated with collecting human preference annotations. These gaps lead to the central question of this work:

Is it possible to achieve superior preference alignment while ensuring thorough privacy protection and maintaining compatibility with emerging preference alignment methodologies?

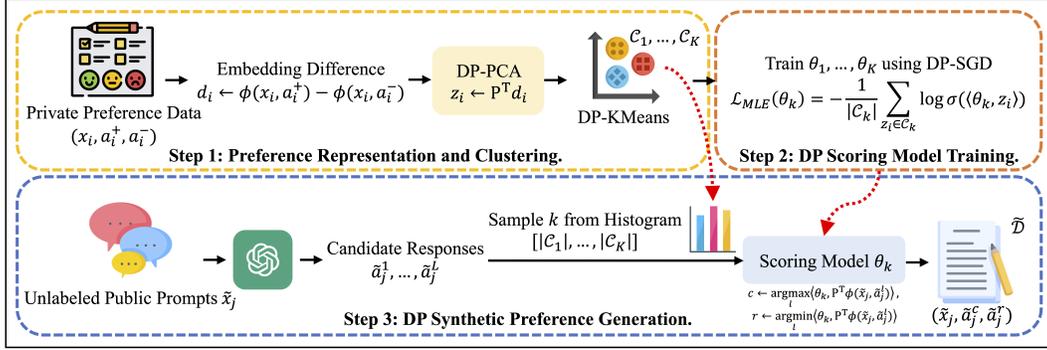


Figure 1: Overview of DPPrefSyn. DPPrefSyn generates DP synthetic preference data through three steps: (1) representing preference samples as embedding differences and clustering them via DP-PCA and DP-KMeans, (2) training DP scoring models on each cluster using DP-SGD, and (3) generating synthetic preference samples from public prompts protected by the DP-protected distribution of scoring models.

In this work, we provide an affirmative answer to this question. Our main contributions are summarized as follows:

- We introduce a new algorithm DPPrefSyn (Figure 1) to generate DP synthetic preference data for LLM alignment. DPPrefSyn features the following elements in its design: 1) *Modeling preference heterogeneity via clustering*. We represent each preference sample using the embedding difference between concatenated prompt-preferred and prompt-dispreferred pairs, and apply clustering to group samples with similar preference patterns. This enables learning per-cluster scoring models that better reflect diverse human preferences. 2) *Improving efficiency through DP-PCA*. To address sample inefficiency and clustering instability in high-dimensional spaces, we apply DP-PCA to project data into a lower-dimensional space while preserving key preference signals under DP. 3) *Saving privacy budgets by using public prompts*. We use public prompts to generate candidate responses and apply private scoring models to select preference pairs. This allows us to focus the privacy budget entirely on modeling private preferences, improving privacy efficiency.
- We conduct a rigorous privacy analysis of DPPrefSyn and confirm it follows (ϵ, δ) -DP. In our analysis, we first allocate a privacy budget ϵ_0 to DP-PCA (Amin et al., 2019) for dimensionality reduction, and ϵ_1 to DP-KMeans (Su et al., 2016) for clustering. The remaining budget is used to train per-cluster scoring models with DP-SGD. We apply the Privacy Random Variable (PRV) accountant (Gopi et al., 2021) to tightly compose the DP-SGD privacy cost, ensuring that DPPrefSyn satisfies (ϵ, δ) -DP. Thanks to the post-processing property of DP, the resulting DP synthetic dataset can be reused across multiple preference alignment methods and different LLMs.
- We empirically evaluate DPPrefSyn on standard benchmarks, including question answering tasks from OpenAssistant (Köpf et al., 2023) and Anthropic-HH (Bai et al., 2022), as well as the TL;DR summarization task (Stiennon et al., 2020). Our experiments suggest that synthetic preference data generated by DPPrefSyn offers strong privacy guarantees while achieving competitive utility across diverse tasks. For example, with $\epsilon = 2$, fine-tuning the Pythia-2.8B model using DPO on our synthetic data achieves a GPT-4o win rate of 56.48% on Anthropic-HH, significantly outperforming fine-tuning on real data (37.02%). These results show that our DP synthetic preference data can enable effective preference alignment of LLMs with strong privacy guarantees. Moreover, DPPrefSyn can be applied to various preference optimization methods (e.g., DPO and RLHF) and remains effective under strong privacy constraints, overcoming the limitations of other privacy-preserving post-training approaches that are tailored to specific methods.

2 RELATED WORK

Differentially Private Alignment of LLMs. In the context of privacy-preserving preference alignment of LLMs, Wu et al. (2023a) first introduce a DP framework to align LLMs with reinforcement learning by adapting the PPO algorithm to the DP setting. However, their approach is *limited to PPO*, while preference alignment continues to evolve with new methods. Chowdhury et al. (2024) study the problem of reward estimation from preference-based feedback, using the notion of label-DP (Ghazi et al., 2021) to protect the privacy of human annotators. Zhang et al. (2025) propose

108 AUP-RLHF, a user-level label-DP framework for RLHF. While both methods are effective at safe-
 109 guarding preference labels, they *do not address the sensitivity of prompts or responses themselves*.
 110 Besides, such methods are usually *constrained by the limited private user preference data*, and may
 111 not achieve high-quality preference alignment, as human preference annotations are costly to obtain.
 112 Yu et al. (2024) focus on the privacy risks of using sensitive user *instructions* in LLM alignment and
 113 propose to generate differentially private synthetic *instructions* to replace real ones during data an-
 114 notation and model fine-tuning.

115 Beyond preference alignment, several studies focus on private fine-tuning of LLMs using general
 116 labeled data, rather than preference comparisons (Yu et al., 2021; Li et al., 2021; Chen et al., 2024;
 117 Tang et al., 2024; Zhang et al., 2023; He et al., 2022), and some recent work focuses on protecting
 118 the privacy of in-context prompts for in-context learning (Duan et al., 2023; Tang et al., 2023; Gao
 119 et al., 2024; Hong et al., 2023; Wu et al., 2023b).

120 **Differentially Private Synthetic Text Generation.** Our work falls within the broader scope of DP
 121 synthetic text generation. In this area, two representative lines of work have emerged. The first
 122 line fine-tunes a language model on private data under DP, and then uses the fine-tuned model to
 123 generate synthetic text (Yue et al., 2023; Mattern et al., 2022; Miresghallah et al., 2023; Carranza
 124 et al., 2023; Yu et al., 2024; Wang et al., 2024; Ochs & Habernal, 2024; Tan et al., 2025; Carranza
 125 et al., 2024). For example, Yu et al. (2024) apply LoRA fine-tuning to LLaMA-7B and LLaMA-13B
 126 models on private instruction data under DP for supervised instruction tuning.

127 Since DP fine-tuning can be expensive and sometimes infeasible (especially for non-public language
 128 models), a recent line of work explores DP synthetic data generation by prompting LLM APIs (Lin
 129 et al., 2023; Xie et al., 2024; Wu et al., 2024; Hou et al., 2024). These approaches typically generate
 130 synthetic samples using LLM API access and then select outputs that are similar to the private data
 131 in a privacy-preserving manner. For instance, Xie et al. (2024) propose the Aug-PE algorithm, which
 132 begins with generating initial synthetic samples using LLM APIs and human-crafted prompts, then
 133 iteratively refines them by selecting examples close to private data in the embedding space using
 134 a DP nearest neighbor histogram. In this work, we follow the general approach in the second line
 135 of work but focus on the preference alignment domain. We train linear models on clustered private
 136 preference data under DP, and use these models to score and select preferred and less preferred
 137 responses from LLM-generated candidates. To the best of our knowledge, this is the first work to
 138 generate a DP synthetic preference dataset for LLM alignment.

139 **Diversity in Human Preferences.** Recent works highlight that human preferences are inherently
 140 diverse (Denton et al., 2021; Aroyo et al., 2023a;b; Chakraborty et al., 2024). The key factors
 141 include sociodemographic differences (e.g., race, gender, age), personal beliefs and biases, varying
 142 levels of domain expertise, and the inherent ambiguity of natural language (Sandri et al., 2023;
 143 Vogels, 2021). Motivated by this, recent studies have explored personalizing preference alignment
 144 to better reflect the values of different user groups (Lee et al., 2024b; Poddar et al., 2024; Singh et al.,
 145 2025). Lee et al. (2024b) train an ensemble (a single neural network with multiple prediction heads),
 146 and dynamically re-weight them at test time using a small set of labeled examples from the target
 147 distribution. Poddar et al. (2024) model user-specific latent variables and learn reward models and
 148 policies conditioned on the latents. Singh et al. (2025) propose a meta-learning framework where
 149 an LLM adapts to individual users using a small number of preference examples. In this work,
 150 we generate DP synthetic preference data that *captures the underlying diversity of human feedback*,
 151 leading to better alignment performance than ignoring such diversity (as shown in Section 5.3).

152 3 PROBLEM DEFINITION, THREAT MODEL, AND NOTATIONS

153
 154 **Preference Alignment of LLMs.** We denote an LLM by a policy π that maps a prompt $x \in \mathcal{X}$ to
 155 a distribution over possible responses in \mathcal{A} . In a typical preference alignment pipeline, a pretrained
 156 LLM π_0 is first fine-tuned using supervised learning (SFT) to adapt it to a downstream task or desired
 157 output style, resulting in π_{SFT} . To further align π_{SFT} with human preferences, a preference dataset
 158 $\mathcal{D} = \{(x_i, a_i^+, a_i^-)\}_{i=1}^n$ is collected. Each sample consists of a prompt $x_i \in \mathcal{X}$, and two responses
 159 $a_i^+, a_i^- \in \mathcal{A}$ generated by a language model, where a_i^+ is labeled as the preferred response and a_i^-
 160 as the less preferred one by human annotators. This preference data is then used to fine-tune π_{SFT} via
 161 methods such as reward modeling and reinforcement learning (e.g., RLHF), or direct optimization
 approaches like DPO. Following Ouyang et al. (2022); Zhu et al. (2023); Rafailov et al. (2024); Liu

et al. (2023), we assume that there exists a ground truth reward function $r^*(x, a) : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$ and the human preference satisfies the Bradley-Terry model (Bradley & Terry, 1952):

$$\mathbb{P}[a_i^+ \succ a_i^- | x_i] = \frac{\exp(r^*(x_i, a_i^+))}{\exp(r^*(x_i, a_i^+)) + \exp(r^*(x_i, a_i^-))} = \sigma(r^*(x_i, a_i^+) - r^*(x_i, a_i^-)),$$

where $a_i^+ \succ a_i^-$ means a_i^+ is preferred to a_i^- , and $\sigma(z) = 1/(1 + e^{-z})$ is the sigmoid function.

In this work, we consider a linear reward model $r_\theta(x, a) = \langle \theta, \phi(x, a) \rangle$, where $\phi : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}^d$ is a known mapping and $\theta \in \mathbb{R}^d$ is a learnable parameter (Saha et al., 2023; Kong & Yang, 2022; Zhu et al., 2023). With this model, we can equivalently write:

$$\mathbb{P}[a_i^+ \succ a_i^- | x_i] = \sigma(\langle \theta, \phi(x_i, a_i^+) - \phi(x_i, a_i^-) \rangle).$$

The most common algorithm for training such reward models in RLHF is maximum likelihood estimation (MLE) (Christiano et al., 2017; Ouyang et al., 2022; Zhu et al., 2023). Given the preference dataset \mathcal{D} , MLE aims at minimizing the negative log likelihood:

$$\mathcal{L}_{\text{MLE}}(\theta) = -\frac{1}{n} \sum_{i=1}^n \log \sigma(\langle \theta, \phi(x_i, a_i^+) - \phi(x_i, a_i^-) \rangle).$$

Differentially Private Preference Data Synthesis. We aim to protect the privacy of the preference dataset $\mathcal{D}_{\text{priv}} = \{(x_i, a_i^+, a_i^-)\}_{i=1}^n$ against an adversary who attempts to access or infer private information about individual prompts, responses, or preference labels. To achieve this, we propose to generate a DP synthetic preference dataset $\tilde{\mathcal{D}} = \{(\tilde{x}_j, \tilde{a}_j^+, \tilde{a}_j^-)\}_{j=1}^m$ that preserves critical preference information in $\mathcal{D}_{\text{priv}}$. This synthetic dataset can be used to fine-tune LLMs using preference alignment methods like DPO or RLHF. Thanks to the post-processing property of DP, $\tilde{\mathcal{D}}$ can be reused across various preference alignment methods and LLMs without incurring additional privacy cost.

We formally define (ϵ, δ) -differential privacy $((\epsilon, \delta)$ -DP) as follows.

Definition 1 $((\epsilon, \delta)$ -Differential Privacy (Dwork et al., 2006a)). A randomized algorithm \mathcal{A} is (ϵ, δ) -differentially private if for any two neighboring inputs \mathcal{D} and \mathcal{D}' that differ by a single entry and any set \mathcal{S} of possible outputs: $\mathbb{P}[\mathcal{A}(\mathcal{D}) \in \mathcal{S}] \leq e^\epsilon \mathbb{P}[\mathcal{A}(\mathcal{D}') \in \mathcal{S}] + \delta$.

4 PROPOSED METHOD

Our framework for preference alignment with DP guarantees consists of two phases:

Phase 1: Generate a DP synthetic dataset $\tilde{\mathcal{D}} = \{(\tilde{x}_j, \tilde{a}_j^+, \tilde{a}_j^-)\}_{j=1}^m$ from the private dataset $\mathcal{D}_{\text{priv}}$;

Phase 2: Use this DP synthetic dataset to fine-tune an LLM using preference alignment algorithms.

The post-processing property of DP ensures that the second phase incurs no additional privacy cost. In the following, we focus on the algorithm that implements the first phase.

4.1 INTUITION BEHIND ALGORITHM DESIGN

To motivate our design, we highlight the key challenges in synthesizing DP preference data. First, human preferences in the private dataset are not homogeneous (Chakraborty et al., 2024; Bakker et al., 2022; Kovač et al., 2023; Jang et al., 2023; Rame et al., 2023; Ji et al., 2023). Different annotators may prioritize different aspects of a response, such as factual accuracy, politeness, creativity, or clarity, resulting in diverse and sometimes conflicting signals in the data. Ignoring this diversity in the underlying preferences can result in synthetic data that fails to reflect the true distribution of $\mathcal{D}_{\text{priv}}$. To address this, we first represent each sample in the private dataset using the embedding difference between the preferred and less preferred responses, conditioned on the same prompt, i.e., $\phi(x_i, a_i^+) - \phi(x_i, a_i^-)$, where ϕ is an embedding model. These vectors capture the implicit direction of preference expressed in each comparison. We then cluster them into multiple groups with similar preference patterns under DP. Within each cluster, we train a private scoring model to guide the selection of synthetic preference pairs that reflect the cluster-specific preference structure.

Second, the embeddings of (prompt, response) pairs $\phi(x, a)$ are intentionally high-dimensional to capture rich semantic and stylistic features. Since training a separate scoring model for each cluster

reduces the effective sample size available to each model, learning in such a high-dimensional space can be sample inefficient. To mitigate this, we apply differentially private PCA (Amin et al., 2019; Liu et al., 2022) to reduce the dimensionality of the representations before clustering and training. This improves sample efficiency and clustering stability while preserving the core preference signals.

Third, synthesizing prompts, preferred responses and less preferred responses simultaneously can quickly exhaust the privacy budget and degrade utility. To address this, we choose to use public prompts to avoid spending privacy budget on synthesizing DP prompts, allowing us to allocate the entire budget to modeling preferences over responses. For each public prompt, we generate multiple candidate responses using an LLM and apply a distribution of scoring functions, which is trained from private preference data under DP, to construct preference pairs. While public prompts may differ in distribution from private ones, our experiments show that, despite this distribution shift in prompts, the synthetic preference data preserves the statistical preference distribution in the private preference dataset and remains effective for downstream preference alignment tasks.

4.2 DPPREFSYN ALGORITHM

We now introduce the proposed DPPrefSyn algorithm (Algorithm 1), which generates a DP synthetic preference dataset from the private dataset $\mathcal{D}_{\text{priv}}$. DPPrefSyn consists of 3 main steps:

Step 1: Preference Representation and Clustering. We encode each prompt-response pair in $\mathcal{D}_{\text{priv}}$ using a public sentence embedding model ϕ . Specifically, for each sample (x_i, a_i^+, a_i^-) , we first concatenate the prompt with each response to form two texts: $[x_i; a_i^+]$ and $[x_i; a_i^-]$. We then compute their embeddings via ϕ , and define their difference as $d_i \leftarrow \phi(x_i, a_i^+) - \phi(x_i, a_i^-)$ (Line 3).

To reduce dimensionality privately, we apply DP-PCA (Amin et al., 2019) to $\{d_i\}_{i=1}^n$, obtaining a projection matrix $\mathbf{P} \in \mathbb{R}^{d \times p}$ under ε_0 -DP (Line 5). Roughly speaking, DP-PCA (Amin et al., 2019) approximates the eigendecomposition of the data covariance matrix by estimating the collections of eigenvalues and eigenvectors separately in a DP manner. Each d_i is then projected to a lower-dimensional space as $z_i \leftarrow \mathbf{P}^\top d_i$ (Line 6). We apply DP-KMeans clustering (Su et al., 2016) on $\{z_i\}_{i=1}^n$ to group samples with similar preference patterns under ε_1 -DP, forming K clusters $\mathcal{C}_1, \dots, \mathcal{C}_K$ in Line 7. Roughly, DP-KMeans (Su et al., 2016) is a DP version of the Lloyd algorithm.

Step 2: DP Scoring Model Training. For each cluster \mathcal{C}_k , we train a linear scoring model $\theta_k \in \mathbb{R}^p$ that captures preference patterns by minimizing the negative log-likelihood (Line 9):

$$\mathcal{L}_{\text{MLE}}(\theta_k) = -\frac{1}{|\mathcal{C}_k|} \sum_{z_i \in \mathcal{C}_k} \log \sigma(\langle \theta_k, z_i \rangle), \quad (1)$$

where z_i is the DP-PCA projected embedding difference from Step 1. Each scoring model is trained with the DP-SGD algorithm (a DP variant of SGD) (Abadi et al., 2016) using noise multiplier σ_2 .

Step 3: DP Synthetic Preference Generation. We compute a histogram $\mathbf{h} = [|\mathcal{C}_1|, \dots, |\mathcal{C}_K|]$ representing the number of samples in each cluster (Line 11). We normalize it to get a probability distribution \mathbf{p} over the clusters, i.e., $\mathbf{p} \leftarrow \mathbf{h}/|\mathcal{D}_{\text{priv}}|$ (Line 12).

Next, for each public prompt \tilde{x}_j , we use a relatively high temperature to prompt an LLM to generate L diverse candidate responses $\tilde{a}_j^1, \dots, \tilde{a}_j^L$ (Line 14). We then sample a cluster index $k \sim \mathbf{p}$ according to the DP histogram (Line 15) and evaluate the responses $\tilde{a}_j^1, \dots, \tilde{a}_j^L$ using the scoring model θ_k associated with the sampled cluster k . To do this, we first compute the embeddings $\phi(\tilde{x}_j, \tilde{a}_j^l)$ for $l = 1, \dots, L$, and project them into the DP-PCA subspace using \mathbf{P} . The preference score for \tilde{a}_j^l is then computed as $\langle \theta_k, \mathbf{P}^\top \phi(\tilde{x}_j, \tilde{a}_j^l) \rangle$. We select the highest- and lowest-scoring responses as the preferred and less preferred responses, respectively: $c \leftarrow \arg \max_{l \in [L]} \langle \theta_k, \mathbf{P}^\top \phi(\tilde{x}_j, \tilde{a}_j^l) \rangle$, $r \leftarrow \arg \min_{l \in [L]} \langle \theta_k, \mathbf{P}^\top \phi(\tilde{x}_j, \tilde{a}_j^l) \rangle$ (Line 16). If the score gap is too small (e.g., < 0.5), we discard the sample to ensure preference quality. Otherwise, we add $(\tilde{x}_j, \tilde{a}_j^c, \tilde{a}_j^r)$ to $\tilde{\mathcal{D}}$. This process is repeated for all public prompts, producing a DP synthetic preference dataset $\tilde{\mathcal{D}}$.

Privacy Analysis of DPPrefSyn. We allocate a privacy budget ε_0 to DP-PCA (Line 5), and ε_1 to DP-KMeans (Line 7), with the remaining budget used to train per-cluster scoring models using DP-SGD with noise multiplier σ_2 (Line 9). For training DP scoring models, we apply parallel composition of DP: since the clusters are disjoint, modifying one data point in the training data

Algorithm 1 DPPrefSyn

```

270 1: Input: Private dataset  $\mathcal{D}_{\text{priv}} = \{(x_i, a_i^+, a_i^-)\}_{i=1}^n$ , public prompts  $\{\tilde{x}_j\}_{j=1}^m$ , embedding model
271  $\phi$ , number of clusters  $K$ , an LLM  $\text{LLM}(\cdot)$ , DP parameters  $\varepsilon_0, \varepsilon_1, \sigma_2$ .
272 2: for each  $(x_i, a_i^+, a_i^-)$  in  $\mathcal{D}_{\text{priv}}$  do
273 3:    $d_i \leftarrow \phi(x_i, a_i^+) - \phi(x_i, a_i^-)$ , where  $d_i \in \mathbb{R}^d$ 
274 4: end for
275 5:  $\mathbf{P} \leftarrow \text{DP-PCA}(\{d_i\}_{i=1}^n, \varepsilon_0)$ , where  $\mathbf{P} \in \mathbb{R}^{d \times p}$ 
276 6:  $\{z_i\}_{i=1}^n \leftarrow \{\mathbf{P}^\top d_i\}_{i=1}^n$ 
277 7:  $\mathcal{C}_1, \dots, \mathcal{C}_K \leftarrow \text{DP-KMeans}(\{z_i\}_{i=1}^n, K, \varepsilon_1)$ 
278 8: for each cluster  $\mathcal{C}_k$  do
279 9:   Train a linear scoring model  $\theta_k$  on  $\mathcal{C}_k$  using DP-SGD with noise multiplier  $\sigma_2$  (Equation (1))
280 10: end for
281 11:  $\mathbf{h} \leftarrow [|\mathcal{C}_1|, \dots, |\mathcal{C}_K|]$ 
282 12:  $\mathbf{p} \leftarrow \mathbf{h} / |\mathcal{D}_{\text{priv}}|$ 
283 13: for each public prompt  $\tilde{x}_j$  do
284 14:   Generate candidate responses  $\tilde{a}_j^1, \dots, \tilde{a}_j^L \leftarrow \text{LLM}(\tilde{x}_j)$ 
285 15:   Sample cluster index  $k \sim \mathbf{p}$ 
286 16:    $c \leftarrow \arg \max_{l \in [L]} \langle \theta_k, \mathbf{P}^\top \phi(\tilde{x}_j, \tilde{a}_j^l) \rangle$ ,  $r \leftarrow \arg \min_{l \in [L]} \langle \theta_k, \mathbf{P}^\top \phi(\tilde{x}_j, \tilde{a}_j^l) \rangle$ 
287 17:   Add  $(\tilde{x}_j, \tilde{a}_j^c, \tilde{a}_j^r)$  to  $\tilde{\mathcal{D}}$ 
288 18: end for
289 19: return  $\tilde{\mathcal{D}}$ .

```

affects only one cluster, so the total privacy cost is bounded by the cost of training on the smallest cluster. We ensure a known lower bound on the cluster size by discarding clusters with too few samples. We use the Privacy Random Variable (PRV) accountant (Gopi et al., 2021) to compose the privacy costs of each DP-SGD step, ensuring that it satisfies $(\varepsilon - \varepsilon_0 - \varepsilon_1, \delta)$ -DP. All other steps involve only public data or post-processing, and incur no additional privacy cost. In our experiments, we set $\varepsilon_0 = \varepsilon_1 = 0.125\varepsilon$, and then select the noise multiplier σ_2 such that the overall privacy guarantee satisfies (ε, δ) -DP. We provide a full proof and code for privacy accounting in Appendix B.

Discussion on Prompt Distribution Shift. DPPrefSyn remains effective even when the prompt distribution is different in private and public datasets. We note that the diverse user preference is captured through θ_k , the parameter in the linear scoring model, instead of $\phi(x, a)$, the feature vector associate with the prompt-response pair. To make this intuitive, consider that a human user may interact with an LLM through diverse prompts, though their underlying preference does not vary. Therefore, as long as we ensure the distribution of θ_k resembles that in the private dataset, the resulting synthetic preferences will still reflect the private preference patterns. This is achieved by randomly sampling a cluster index k according to the DP histogram (Line 15) for each public prompt, and then select the synthetic response pairs accordingly. Therefore, even if the prompt distribution may differ in private and public datasets, our algorithm can still preserve the preference patterns captured in the private data under DP constraint.

5 EXPERIMENTS

Datasets. We study two types of tasks: question answering and summarization. For question answering, we use the OpenAssistant dataset (Köpf et al., 2023), which contains assistant-style conversations, and the Anthropic-HH dataset (Bai et al., 2022), which provides human preference comparisons focused on helpfulness and harmlessness. For summarization, we use the TL;DR dataset (Stiennon et al., 2020), which contains annotations of human preference on pairs of summaries. To simulate public data, we use prompts from Alpaca (Taori et al., 2023) for the OpenAssistant QA task, SafeRLHF (Ji et al., 2024) for the Anthropic-HH QA task, and XSum (Narayan et al., 2018) for the TL;DR summarization task. Additional dataset details are available in Appendix C.2.

Setup for Algorithm 1. By default, we use BAAI/bge-large-en-v1.5 (Xiao et al., 2023) as the embedding model ϕ to encode prompt-response pairs. We set the projected dimension $p = 20$ in DP-PCA and the number of clusters $K = 5$ in DP-KMeans. For DP-SGD, we use a learning rate of 0.1, a batch size of 4, 4 training epochs, and a gradient clipping norm of 1.0. To generate candidate

Table 1: GPT-4o win rates on test sets with the downstream Pythia-2.8B model fine-tuned by SFT and DPO. $\epsilon = 0$ denotes the base (fully private, non-finetuned) LLM. Compared with DP-FT, DPPrefSyn achieves higher win rates under privacy levels $\epsilon = 1, 2, 4, 8$, even outperforming the utility of non-private fine-tuning baseline (DP-FT with $\epsilon = \infty$). Results show the mean and standard deviation over 5 runs using different random seeds.

Task	$\epsilon = 0$	Method	Data Type	Pref. Alignment	$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = \infty$
OpenAssistant	2.11	DP-FT	Original	SFT	3.96 _{0.20}	4.29 _{0.21}	4.38 _{0.23}	4.46 _{0.18}	4.49 _{0.30}	4.75 _{0.27}
				SFT+DPO	5.65 _{0.32}	6.04 _{0.58}	6.18 _{0.83}	6.12 _{0.63}	6.32 _{0.71}	8.20 _{0.62}
Anthropic-HH	12.14	DPPrefSyn	Synthetic	SFT	8.96 _{0.39}	9.07 _{0.32}	9.10 _{0.51}	9.33 _{0.56}	9.63 _{0.38}	9.80 _{0.58}
				SFT+DPO	9.86 _{0.59}	10.31 _{0.54}	11.04 _{0.92}	11.85 _{0.99}	12.05 _{0.82}	12.36 _{0.55}
TL;DR	11.64	DP-FT	Original	SFT	29.77 _{0.76}	31.10 _{0.72}	31.53 _{0.77}	31.29 _{0.65}	31.37 _{0.67}	31.98 _{0.56}
				SFT+DPO	35.00 _{0.49}	36.27 _{0.63}	37.02 _{0.72}	36.74 _{0.90}	36.94 _{1.06}	38.72 _{0.56}
TL;DR	11.64	DPPrefSyn	Synthetic	SFT	54.90 _{0.66}	55.19 _{0.91}	55.21 _{0.82}	54.95 _{1.10}	55.63 _{0.86}	55.95 _{0.70}
				SFT+DPO	55.08 _{1.04}	55.96 _{0.80}	56.48 _{0.61}	56.51 _{1.93}	56.86 _{1.17}	57.53 _{0.93}
TL;DR	11.64	DP-FT	Original	SFT	21.19 _{1.12}	22.41 _{1.09}	22.57 _{0.65}	22.45 _{0.90}	23.04 _{0.79}	23.82 _{1.26}
				SFT+DPO	59.69 _{0.82}	62.21 _{1.12}	62.02 _{1.67}	63.00 _{0.73}	63.06 _{0.84}	64.52 _{1.79}
TL;DR	11.64	DPPrefSyn	Synthetic	SFT	61.92 _{1.06}	62.14 _{1.61}	62.49 _{1.48}	62.72 _{1.66}	62.29 _{1.14}	62.55 _{1.23}
				SFT+DPO	67.50 _{0.79}	68.07 _{1.95}	68.54 _{2.76}	68.56 _{1.59}	71.01 _{2.17}	72.20 _{1.08}

responses for public prompts, we use the instruction-finetuned LLaMA-7B-chat model (Touvron et al., 2023) with a temperature of 0.9. We set the number of candidates per prompt to $L = 5$. We consider overall privacy budgets $\epsilon = 1, 2, 4, 8$ and set $\delta = 1/|D_{\text{priv}}|$. We study the effect of different generator models in Table 3, embedding models and hyperparameters in Section 5.3.

Evaluation. For downstream preference alignment, we fine-tune the Pythia-2.8B model (Biderman et al., 2023). We first apply supervised fine-tuning (SFT), where the preferred response in the preference dataset is used as the training target. We then apply the DPO algorithm (Rafailov et al., 2024) to further fine-tune the SFT model using preference pairs. Following Rafailov et al. (2024), we measure the win rate of the model-generated responses against the preferred responses in the test set using the GPT-4o model. We study more types of downstream models, and different preference alignment methods as ablation study in Section 5.2.

Baselines. We compare DPPrefSyn with fine-tuning the downstream model on real data using DP-SGD (Abadi et al., 2016), denoted as DP-FT. Note that DPPrefSyn is more reusable and versatile than DP-FT because it generates DP synthetic data that can be used to train different models with different alignment algorithms without incurring additional privacy cost, while DP-FT consumes the privacy budget in training a single privatized model. We also include a fully private baseline $\epsilon = 0$, where we evaluate the base model without any preference fine-tuning.

Details about the setups, hyperparameters, metrics, and baselines are provided in Appendix C.3.

5.1 UNDERSTANDING THE PERFORMANCE OF DPPREFSYN

In this section, we analyze the performance of DPPrefSyn by answering two research questions about its privacy-utility trade-off. Our main results are presented in Table 1.

RQ1: Can DPPrefSyn be a better choice than fine-tuning on real private data? DPPrefSyn consistently achieves a stronger privacy-utility trade-off than DP-FT across all tasks under $\epsilon = 1, 2, 4, 8$, and outperforms the utility of fine-tuning on real private data without DP constraints (DP-FT with $\epsilon = \infty$), while preserving privacy. For instance, on the OpenAssistant QA task with $\epsilon = 2$, our approach achieves 9.10% after SFT and 11.04% after DPO, compared to DP-FT’s 4.38% after SFT and 6.18% after DPO. Remarkably, while still preserving privacy, DPPrefSyn achieves higher utility than the non-private baseline of fine-tuning on real private data (DP-FT with $\epsilon = \infty$), which reaches 4.75% after SFT and 8.20% after DPO. Similarly, on the Anthropic-HH dataset with $\epsilon = 2$, DPPrefSyn achieves a 55.21% win rate after SFT and 56.48% after DPO, compared to 31.53% and 37.02% achieved by DP-FT, respectively. It also outperforms direct fine-tuning on real data (DP-FT with $\epsilon = \infty$), which achieves 31.98% after SFT and 38.72% after DPO.

These results highlight two **key advantages** of our DP synthetic data. **First**, DPPrefSyn leads to significant performance improvement after SFT alone compared to directly applying SFT on private data. For example, on the TL;DR summarization task with $\epsilon = 2$, DPPrefSyn achieves a win rate of 62.49% after SFT, much higher than the 22.57% achieved by SFT on private data. This is mainly because DPPrefSyn generates candidate responses using a high-performance LLM, leading to high-

quality training data that provides more effective supervision during SFT. This advantage holds even with limited public data. For instance, on the OpenAssistant QA task with $\epsilon = 4$ and only 5K public prompts, DPPrefSyn achieves a win rate of 6.80% after SFT, compared to 4.46% for SFT directly on private data (Table 4), which contains around 14K samples, highlighting DPPrefSyn’s ability to handle data scarcity while preserving utility. **Second**, applying DPO with our synthetic data on top of the SFT model leads to further improvements, showing that our synthetic preference pairs capture meaningful distinctions aligned with human preferences. This is enabled by the private scoring functions that capture diverse user preferences while preserving privacy. Compared to directly using private data, the improvement from DPO is smaller in our setting, likely because the SFT model trained on our synthetic data already performs well, leaving less room for optimization. However, for the entire fine-tuning process, the performance gain using our DP synthetic dataset is still significantly higher than that using real private data.

RQ2: How does DPPrefSyn perform across different privacy budget ϵ ? **First**, Table 1 shows that DPPrefSyn generally achieves better performance compared to DP-FT as ϵ increases from 1, 2, 4, 8 to ∞ , indicating good scalability with privacy budget ϵ . **Second**, DPPrefSyn can be more robust to DP noise than DP-FT in some cases. For example, we find that on OpenAssistant with Llama-3-8B as the downstream model, DPO performance for DP-FT drops from 59.13% to 45.79% when ϵ decreases from ∞ to 4, while DPPrefSyn drops only slightly from 61.52% to 60.25% (see Table 3 in Section 5.2 for details). The reason could be that DP-FT perturbs model parameters directly through DP-SGD, whereas DPPrefSyn adds noise during data synthesis, which better preserves utility.

5.2 UNDERSTANDING THE PROPERTIES OF DPPREFSYN

In this section, we study properties of DPPrefSyn, including its generalizability across different preference alignment methods and LLMs, its performance under varying scales of public data, and its robustness to empirical privacy attacks.

RQ3: Is DPPrefSyn effective across different preference alignment methods? **DPPrefSyn supports various preference alignment methods and remains effective under strong DP constraints.** In Table 2, we evaluate DPPrefSyn with SFT, DPO, and PPO. For PPO, we train K reward models (one for each cluster) to capture diverse preference patterns. During PPO training, we use the average output of the K reward models as the reward and optimize the policy. We find that DPPrefSyn consistently achieves better utility for SFT, DPO, and PPO, compared with fine-tuning on private data, while preserving strong privacy guarantees.

RQ4: Can more powerful downstream LLMs benefit from synthetic data generated by DPPrefSyn? **DPPrefSyn can be effective across a wide range of downstream LLMs.** In Table 3, we conduct experiments on different downstream models, including Qwen-3-4B, Gemma-2-2B, and Llama-3-8B, using Llama-3-8B-Instruct and Qwen-3-4B-Instruct-2507 as generator models. We find that DPPrefSyn can outperform fine-tuning on real data, while ensuring privacy across different model families after DPO. For example, using Qwen-3-4B-Instruct-2507 as the generator, DPPrefSyn achieves an 86.52% win rate on Qwen-3-4B with $\epsilon = 4$, which significantly outperforms non-private fine-tuning on real data (59.41%). We also observe that DPPrefSyn’s performance improves with stronger generator models, with Qwen-3-4B-Instruct-2507 outperforming Llama-3-8B-Instruct.

Table 2: DPPrefSyn achieves higher win rates than fine-tuning on private data across SFT, DPO, and PPO, while providing strong privacy guarantees on OpenAssistant.

Pref. Alignment	$\epsilon = 0$	DPPrefSyn ($\epsilon = 4$)	DP-FT ($\epsilon = \infty$)
SFT		9.33_{0.56}	4.75 _{0.27}
SFT+DPO	2.11	11.85_{0.99}	8.20 _{0.62}
SFT+PPO		10.11_{0.61}	7.08 _{0.78}

Table 3: DPPrefSyn achieves higher win rates than fine-tuning on private data across different downstream LLMs after DPO, while providing strong privacy guarantees on OpenAssistant. DPPrefSyn can also benefit from more powerful generator LLMs.

Downstream LLM	$\epsilon = 0$	Method	Generator LLM	$\epsilon = 4$	$\epsilon = \infty$
Qwen-3-4B	48.17	DP-FT	-	54.92	59.41
		DPPrefSyn	Qwen-3-4B-Instruct Llama-3-8B-Instruct	86.52 51.69	87.08 52.11
		DP-FT	-	30.06	33.85
Gemma-2-2B	7.30	DPPrefSyn	Qwen-3-4B-Instruct Llama-3-8B-Instruct	36.24 36.01	37.08 36.66
		DP-FT	-	45.79	59.13
		DPPrefSyn	Qwen-3-4B-Instruct Llama-3-8B-Instruct	60.25 42.28	61.52 44.10

RQ5: Can DPPrefSyn be effective with varying scales of public data? We study the effect of public data size m in Table 4. The private dataset (OpenAssistant) contains 14K preference annotations, while the full public dataset (Alpaca) contains 52K prompts. We randomly sample subsets of the public prompts at different sizes ($m = 5K, 10K, 52K$) to generate synthetic preference data using DPPrefSyn. We find that even with only 5K public prompts, DPPrefSyn achieves a win rate of 6.80% after SFT and 9.44% after DPO under $\epsilon = 4$, outperforming the DP-FT baseline, which achieves 4.46% and 6.12%, respectively. This shows that **DPPrefSyn remains effective even when the available public data is limited**, thanks to the high-quality responses generated by the LLMs and the private scoring functions that accurately capture user preferences. Besides, **as the public data size increases, the performance improvement over the baseline becomes more prominent**, indicating that DPPrefSyn can effectively leverage the vast amounts of public data for preference alignment, further mitigating the data scarcity issue in private preference data.

RQ6: Can DPPrefSyn effectively mitigate empirical privacy attacks? In Table 5, we perform membership inference attacks (MIAs) (Shokri et al., 2017) against the DPO fine-tuned downstream models on OpenAssistant and report the AUC score. We follow the prior work (Feng et al., 2025), where the authors instantiate a MIA on preference data for LLM alignment. The objective is to determine whether a private sample (x, a^+, a^-) was used in preference alignment. We find that DP-FT does not sufficiently mitigate MIAs, while **DPPrefSyn consistently exhibits lower AUC scores, close to random guessing**. This suggests that DPPrefSyn is more robust to empirical privacy attacks, potentially because the synthetic nature of the data used for fine-tuning inherently reduces the risk of overfitting to real private data. We defer the details to Appendix C.4.

5.3 VALIDATING THE DESIGN OF DPPREFSYN

We validate the design of DPPrefSyn by studying its algorithmic components in this section.

RQ7: How does modeling diverse human preferences impact DPPrefSyn’s performance? We evaluate this by removing either DP-KMeans or DP-PCA, and varying the number of clusters K , using OpenAssistant with $\epsilon = 4$. **(i) Removing either DP-KMeans or DP-PCA.** Figure 2 shows that removing either DP-KMeans or DP-PCA leads to a performance drop. Without DP-KMeans, all samples are assigned to a single cluster. Removing DP-PCA leads to ineffective clustering and sample-inefficient learning in a high-dimensional space. Using multiple clusters with both components improves performance over using a single cluster. **(ii) Varying the number of clusters K .** In Figure 3, we evaluate DPPrefSyn with $K = 1, 3, 5, 8$ clusters. As K increases, performance first improves at $K = 3$ and then remains relatively stable at $K = 5, 8$. This suggests that DPPrefSyn is robust to K , and moderate values (e.g., $K = 3, 5$) provide a good trade-off between diversity and data efficiency. Overall, these results show the benefit of capturing diverse preferences using multiple clusters in DPPrefSyn.

RQ8: How do different embedding models or hyperparameters affect DPPrefSyn’s performance? We perform ablation studies on the embedding model ϕ , the DP-PCA dimension p , the number of candidates per prompt L , and the filter threshold, with detailed results in Appendix C.5. Across different sentence encoders, including bge-large-en-v1.5 (Xiao et al., 2023), all-mpnet-base-v2 (Song et al., 2020), and all-distilroberta-v1 (Sanh et al.,

Table 4: DPPrefSyn achieves higher win rates than fine-tuning on private data with varying public data size m .

Data Type (Size)	DPPrefSyn ($\epsilon = 4$)			DP-FT ($\epsilon = 4$)	DP-FT ($\epsilon = \infty$)
	Synthetic ($m = 5,000$)	Synthetic ($m = 10,000$)	Synthetic ($m = 52,000$)	Original (14,167)	Original (14,167)
SFT	6.80 _{0.66}	7.30 _{0.48}	9.33 _{0.56}	4.46 _{0.18}	4.75 _{0.27}
SFT+DPO	9.44 _{0.45}	9.55 _{0.92}	11.85 _{0.99}	6.12 _{0.63}	8.20 _{0.62}

Table 5: DPPrefSyn exhibits lower AUC scores against MIAs (Feng et al., 2025) on OpenAssistant, close to random guessing.

Downstream LLM	Method	Generator LLM	$\epsilon = 4$	$\epsilon = \infty$
Qwen-3-4B	DP-FT	-	69.50	72.66
	DPPrefSyn	Qwen-3-4B-Instruct	50.92	51.03
		Llama-3-8B-Instruct	49.78	50.57
Gemma-2-2B	DP-FT	-	61.82	63.81
	DPPrefSyn	Qwen-3-4B-Instruct	49.91	51.38
		Llama-3-8B-Instruct	51.03	50.57
Llama-3-8B	DP-FT	-	61.07	68.76
	DPPrefSyn	Qwen-3-4B-Instruct	49.55	50.45
		Llama-3-8B-Instruct	49.39	51.41

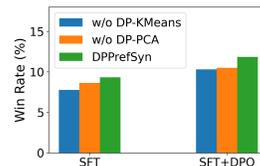


Figure 2: Removing DP-KMeans or DP-PCA.

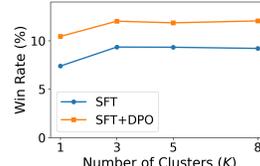


Figure 3: Varying clusters.

2019), we find that DPPrefSyn remains consistently effective. For DP-PCA, $p = 20$ provides the best overall performance after DPO. For the number of candidates per prompt, $L = 5$ serves as a good default choice. For filtering, a moderate threshold of 0.5 slightly outperforms no filtering, while a high threshold of 1 is too aggressive and hurts performance.

We discuss our design choice of using linear scoring models in Appendix C.6, which shows that the linear model can capture preference signals effectively despite its simplicity. In Appendix C.7, we analyze cluster interpretability and show that the clusters are meaningful and interpretable in terms of user preference styles. We provide generated samples using DPPrefSyn in Appendix C.9.

6 CONCLUSION

In this work, we introduced DPPrefSyn, an algorithm for synthesizing DP preference data from a private dataset. By modeling preference diversity via clustering, improving efficiency with DP-PCA, and conserving privacy budgets through public prompts, DPPrefSyn enables effective and private preference data synthesis for preference alignment of LLMs. Our experiments show that DPPrefSyn achieves strong privacy guarantees while outperforming fine-tuning on real data. One limitation is that we only focus on text-based alignment tasks. A potential future direction is to extend our method to support private alignment in multi-modal settings, including image and speech.

540 ETHICS STATEMENT

541
542 This work enables preference alignment of large language models without exposing personal in-
543 formation, which helps balance the benefits of big data with the need to protect individual privacy,
544 promoting ethical data usage and fostering societal trust in an increasingly data-driven world.
545

546 REPRODUCIBILITY STATEMENT

547
548 We provide the full privacy analysis of DPPrefSyn in Appendix B. Detailed experimental setups,
549 including computational resources, datasets, implementation details, and hyperparameter configura-
550 tions, are provided in Appendix C. The reproducible code is included in the Supplementary Material.
551

552 REFERENCES

- 553
554 Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and
555 Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC*
556 *conference on computer and communications security*, pp. 308–318, 2016.
557
- 558 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-
559 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical
560 report. *arXiv preprint arXiv:2303.08774*, 2023.
- 561 Kareem Amin, Travis Dick, Alex Kulesza, Andres Munoz, and Sergei Vassilvitskii. Differentially
562 private covariance estimation. *Advances in Neural Information Processing Systems*, 32, 2019.
563
- 564 Lora Aroyo, Mark Diaz, Christopher Homan, Vinodkumar Prabhakaran, Alex Taylor, and Ding
565 Wang. The reasonable effectiveness of diverse evaluation data. *arXiv preprint arXiv:2301.09406*,
566 2023a.
- 567 Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García,
568 Vinodkumar Prabhakaran, and Ding Wang. Dices dataset: Diversity in conversational ai evalua-
569 tion for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342, 2023b.
570
- 571 Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn
572 Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless
573 assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*,
574 2022.
- 575 Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham,
576 Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning
577 language models to find agreement among humans with diverse preferences. *Advances in Neural*
578 *Information Processing Systems*, 35:38176–38189, 2022.
579
- 580 Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric
581 Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al.
582 Pythia: A suite for analyzing large language models across training and scaling. In *International*
583 *Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.
- 584 Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method
585 of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.
586
- 587 Aldo Carranza, Reza Farahani, Natalia Ponomareva, Alexey Kurakin, Matthew Jagielski, and Milad
588 Nasr. Synthetic query generation for privacy-preserving deep retrieval systems using differentially
589 private language models. In *Proceedings of the 2024 Conference of the North American Chapter*
590 *of the Association for Computational Linguistics: Human Language Technologies (Volume 1:*
591 *Long Papers)*, pp. 3920–3930, 2024.
- 592 Aldo Gael Carranza, Reza Farahani, Natalia Ponomareva, Alex Kurakin, Matthew Jagielski, and
593 Milad Nasr. Privacy-preserving recommender systems with synthetic query generation using
differentially private large language models. *arXiv preprint arXiv:2305.05973*, 2023.

- 594 Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit
595 Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences. In *International
596 Conference on Machine Learning*, pp. 6116–6135. PMLR, 2024.
- 597
598 Tiejin Chen, Longchao Da, Huixue Zhou, Pingzhi Li, Kaixiong Zhou, Tianlong Chen, and Hua
599 Wei. Privacy-preserving fine-tuning of large language models through flatness. *arXiv preprint
600 arXiv:2403.04124*, 2024.
- 601 Sayak Ray Chowdhury, Xingyu Zhou, and Nagarajan Natarajan. Differentially private reward es-
602 timation with preference feedback. In *International Conference on Artificial Intelligence and
603 Statistics*, pp. 4843–4851. PMLR, 2024.
- 604
605 Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep
606 reinforcement learning from human preferences. *Advances in neural information processing sys-
607 tems*, 30, 2017.
- 608 Remi Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. Whose
609 ground truth? accounting for individual and collective identities underlying dataset annotation.
610 *arXiv preprint arXiv:2112.04554*, 2021.
- 611
612 Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. Flocks of stochastic
613 parrots: Differentially private prompt learning for large language models. *Advances in Neural
614 Information Processing Systems*, 36:76852–76871, 2023.
- 615 Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data,
616 ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT
617 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic
618 Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25*, pp. 486–503. Springer,
619 2006a.
- 620
621 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity
622 in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference,
623 TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pp. 265–284. Springer, 2006b.
- 624
625 Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations
626 and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.
- 627
628 Qizhang Feng, Siva Rajesh Kasa, SANTHOSH KUMAR KASA, Hyokun Yun, Choon Hui Teo,
629 and Sravan Babu Bodapati. Exposing privacy gaps: Membership inference attack on preference
630 data for llm alignment. In *International Conference on Artificial Intelligence and Statistics*, pp.
5221–5229. PMLR, 2025.
- 631
632 Fengyu Gao, Ruida Zhou, Tianhao Wang, Cong Shen, and Jing Yang. Data-adaptive differentially
633 private prompt synthesis for in-context learning. *arXiv preprint arXiv:2410.12085*, 2024.
- 634
635 Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning
636 with label differential privacy. *Advances in neural information processing systems*, 34:27131–
27145, 2021.
- 637
638 Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy.
639 *Advances in Neural Information Processing Systems*, 34:11631–11642, 2021.
- 640
641 Jiyan He, Xuechen Li, Da Yu, Huishuai Zhang, Janardhan Kulkarni, Yin Tat Lee, Arturs Backurs,
642 Nenghai Yu, and Jiang Bian. Exploring the limits of differentially private deep learning with
group-wise clipping. *arXiv preprint arXiv:2212.01539*, 2022.
- 643
644 Naoise Holohan, Stefano Braghin, Pól Mac Aonghusa, and Killian Levacher. Diffprivlib: the IBM
645 differential privacy library. *ArXiv e-prints*, 1907.02444 [cs.CR], July 2019.
- 646
647 Junyuan Hong, Jiachen T Wang, Chenhui Zhang, Zhangheng Li, Bo Li, and Zhangyang Wang.
Dp-opt: Make large language model your privacy-preserving prompt engineer. *arXiv preprint
arXiv:2312.03724*, 2023.

- 648 Charlie Hou, Akshat Shrivastava, Hongyuan Zhan, Rylan Conway, Trang Le, Adithya Sagar, Giulia
649 Fanti, and Daniel Lazar. Pre-text: Training language models on private federated data in the age
650 of llms. In *International Conference on Machine Learning*, pp. 19043–19061. PMLR, 2024.
651
- 652 Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer,
653 Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Per-
654 sonalized large language model alignment via post-hoc parameter merging. *arXiv preprint*
655 *arXiv:2310.11564*, 2023.
- 656 Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun,
657 Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via
658 a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–
659 24704, 2023.
660
- 661 Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu,
662 Boxun Li, and Yaodong Yang. Pku-saferllhf: Towards multi-level safety alignment for llms with
663 human preference. *arXiv preprint arXiv:2406.15513*, 2024.
- 664 Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward
665 Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and
666 diversity. In *The Twelfth International Conference on Learning Representations*.
667
- 668 Dingwen Kong and Lin Yang. Provably feedback-efficient reinforcement learning via active reward
669 learning. *Advances in Neural Information Processing Systems*, 35:11063–11078, 2022.
670
- 671 Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith
672 Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant
673 conversations-democratizing large language model alignment. *Advances in Neural Information*
674 *Processing Systems*, 36:47669–47681, 2023.
- 675 Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-
676 Yves Oudeyer. Large language models as superpositions of cultural perspectives. *arXiv preprint*
677 *arXiv:2307.07870*, 2023.
678
- 679 Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu,
680 Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling re-
681 inforcement learning from human feedback with ai feedback. In *International Conference on*
682 *Machine Learning*, pp. 26874–26901. PMLR, 2024a.
- 683 Yoonho Lee, Jonathan Williams, Henrik Marklund, Archit Sharma, Eric Mitchell, Anikait
684 Singh, and Chelsea Finn. Test-time alignment via hypothesis reweighting. *arXiv preprint*
685 *arXiv:2412.08812*, 2024b.
686
- 687 Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, Fanpu Meng, and Yangqiu Song. Multi-
688 step jailbreaking privacy attacks on chatgpt. In *Findings of the Association for Computational*
689 *Linguistics: EMNLP 2023*, pp. 4138–4153, 2023.
- 690 Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be
691 strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021.
692
- 693 Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, Harsha Nori, and Sergey Yekhanin. Differentially
694 private synthetic data via foundation model apis 1: Images. *arXiv preprint arXiv:2305.15560*,
695 2023.
- 696 Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and
697 Jialu Liu. Statistical rejection sampling improves preference optimization. *arXiv preprint*
698 *arXiv:2309.06657*, 2023.
699
- 700 Xiyang Liu, Weihao Kong, Prateek Jain, and Sewoong Oh. Dp-pca: Statistically optimal and dif-
701 ferentially private pca. *Advances in neural information processing systems*, 35:29929–29943,
2022.

- 702 Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan.
703 Differentially private language models for secure data sharing. In *Proceedings of the 2022 Con-*
704 *ference on Empirical Methods in Natural Language Processing*, pp. 4860–4873, 2022.
- 705
- 706 Fatemehsadat Mireshghallah, Yu Su, Tatsunori B Hashimoto, Jason Eisner, and Richard Shin.
707 Privacy-preserving domain adaptation of semantic parsers. In *Proceedings of the 61st Annual*
708 *Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4950–
709 4970, 2023.
- 710 Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don’t give me the details, just the sum-
711 mary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint*
712 *arXiv:1808.08745*, 2018.
- 713
- 714 Sebastian Ochs and Ivan Habernal. Private synthetic text generation with diffusion models. *arXiv*
715 *preprint arXiv:2410.22971*, 2024.
- 716
- 717 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong
718 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to fol-
719 low instructions with human feedback. *Advances in neural information processing systems*, 35:
720 27730–27744, 2022.
- 721 Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing
722 reinforcement learning from human feedback with variational preference learning. *arXiv preprint*
723 *arXiv:2408.10075*, 2024.
- 724
- 725 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea
726 Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances*
727 *in Neural Information Processing Systems*, 36, 2024.
- 728 Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor,
729 Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by in-
730 terpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing*
731 *Systems*, 36:71095–71134, 2023.
- 732
- 733 Aadirupa Saha, Aldo Pacchiano, and Jonathan Lee. Dueling rl: Reinforcement learning with trajec-
734 tory preferences. In *International conference on artificial intelligence and statistics*, pp. 6263–
735 6289. PMLR, 2023.
- 736
- 737 Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Ježek. Why don’t you do it right?
738 analysing annotators’ disagreement in subjective tasks. In *Proceedings of the 17th Conference of*
739 *the European Chapter of the Association for Computational Linguistics*, pp. 2428–2441, 2023.
- 740
- 741 Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of
742 bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- 743
- 744 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
745 Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical
746 reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- 747
- 748 Bo Shen, Jiaxin Zhang, Taihong Chen, Daoguang Zan, Bing Geng, An Fu, Muhan Zeng, Ailun Yu,
749 Jichuan Ji, Jingyang Zhao, et al. Pangu-coder2: Boosting large language models for code with
750 ranking feedback. *arXiv preprint arXiv:2307.14936*, 2023.
- 751
- 752 Michael Shoemate, Andrew Vyrros, Chuck McCallum, Raman Prasad, Philip Durbin, Sílvia
753 Casacuberta Puig, Ethan Cowan, Vicki Xu, Zachary Ratliff, Nicolás Berrios, Alex Whitworth,
754 Michael Eliot, Christian Lebeda, Oren Renard, and Claire McKay Bowen. OpenDP Library,
755 2021. URL <https://github.com/opendp/opendp>.
- 756
- 757 Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference at-
758 tacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*,
759 pp. 3–18. IEEE, 2017.

- 756 Anikait Singh, Sheryl Hsu, Kyle Hsu, Eric Mitchell, Stefano Ermon, Tatsunori Hashimoto, Archit
757 Sharma, and Chelsea Finn. Fspo: Few-shot preference optimization of synthetic preference data
758 in llms elicits effective personalization to real users. *arXiv preprint arXiv:2502.19312*, 2025.
- 759 Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-
760 training for language understanding. *Advances in neural information processing systems*, 33:
761 16857–16867, 2020.
- 762 Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford,
763 Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances
764 in neural information processing systems*, 33:3008–3021, 2020.
- 765 Dong Su, Jianneng Cao, Ninghui Li, Elisa Bertino, and Hongxia Jin. Differentially private k-means
766 clustering. In *Proceedings of the sixth ACM conference on data and application security and
767 privacy*, pp. 26–37, 2016.
- 768 Bowen Tan, Zheng Xu, Eric Xing, Zhiting Hu, and Shanshan Wu. Synthesizing privacy-preserving
769 text data via finetuning without finetuning billion-scale llms. *arXiv preprint arXiv:2503.12347*,
770 2025.
- 771 Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin,
772 Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. Privacy-preserving in-context learning with
773 differentially private few-shot generation. *arXiv preprint arXiv:2309.11765*, 2023.
- 774 Xinyu Tang, Ashwinee Panda, Milad Nasr, Saeed Mahloujifar, and Prateek Mittal. Pri-
775 vate fine-tuning of large language models with zeroth-order optimization. *arXiv preprint
776 arXiv:2401.04343*, 2024.
- 777 Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin,
778 Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-
779 following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, 3(6):7, 2023.
- 780 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée
781 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and
782 efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- 783 Emily A Vogels. *The state of online harassment*, volume 13. Pew Research Center Washington, DC,
784 2021.
- 785 WenHao Wang, Xiaoyu Liang, Rui Ye, Jingyi Chai, Siheng Chen, and Yanfeng Wang. Knowl-
786 edgesg: Privacy-preserving synthetic text generation with knowledge distillation from server. In
787 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp.
788 7677–7695, 2024.
- 789 Fan Wu, Huseyin A Inan, Arturs Backurs, Varun Chandrasekaran, Janardhan Kulkarni, and
790 Robert Sim. Privately aligning language models with reinforcement learning. *arXiv preprint
791 arXiv:2310.16960*, 2023a.
- 792 Shanshan Wu, Zheng Xu, Yanxiang Zhang, Yuanbo Zhang, and Daniel Ramage. Prompt pub-
793 lic large language models to synthesize data for private on-device applications. *arXiv preprint
794 arXiv:2404.04360*, 2024.
- 795 Tong Wu, Ashwinee Panda, Jiachen T Wang, and Prateek Mittal. Privacy-preserving in-context
796 learning for large language models. *arXiv e-prints*, pp. arXiv–2305, 2023b.
- 797 Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to
798 advance general chinese embedding, 2023.
- 799 Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A Inan, Harsha Nori, Hao-
800 tian Jiang, Huishuai Zhang, Yin Tat Lee, et al. Differentially private synthetic data via foundation
801 model apis 2: Text. In *International Conference on Machine Learning*, pp. 54531–54560. PMLR,
802 2024.

- 810 Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang.
811 Iterative preference learning from human feedback: Bridging theory and practice for rlhf under
812 kl-constraint. *arXiv preprint arXiv:2312.11456*, 2023.
- 813
814 Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani
815 Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, et al. Opacus: User-friendly
816 differential privacy library in pytorch. *arXiv preprint arXiv:2109.12298*, 2021.
- 817 Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan
818 Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning
819 of language models. *arXiv preprint arXiv:2110.06500*, 2021.
- 820
821 Da Yu, Peter Kairouz, Sewoong Oh, and Zheng Xu. Privacy-preserving instructions for aligning
822 large language models. In *International Conference on Machine Learning*, pp. 57480–57506.
823 PMLR, 2024.
- 824 Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun,
825 David Levitan, and Robert Sim. Synthetic text generation with differential privacy: A simple and
826 practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational
827 Linguistics (Volume 1: Long Papers)*, pp. 1321–1342, 2023.
- 828 Jiaming Zhang, Mingxi Lei, Meng Ding, Mengdi Li, Zihang Xiang, Difei Xu, Jinhui Xu, and
829 Di Wang. Towards user-level private reinforcement learning with human feedback. *arXiv preprint
830 arXiv:2502.17515*, 2025.
- 831
832 Liang Zhang, Kiran Koshy Thekumparampil, Sewoong Oh, and Niao He. Dpzero:
833 Dimension-independent and differentially private zeroth-order optimization. *arXiv preprint
834 arXiv:2310.09639*, 2023.
- 835 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,
836 Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and
837 chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.
- 838
839 Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feed-
840 back from pairwise or k-wise comparisons. In *International Conference on Machine Learning*,
841 pp. 43037–43067. PMLR, 2023.
- 842 Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul
843 Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv
844 preprint arXiv:1909.08593*, 2019.
- 845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863

864 A THE USE OF LARGE LANGUAGE MODELS

865
866 In the design of our proposed DPPrefSyn, we use LLMs for synthetic data generation by prompting
867 LLM APIs. We also perform preference alignment methods (e.g., RLHF and DPO) on LLMs, as
868 detailed in Section 5.

869 Additionally, LLMs are used during the preparation of this paper to improve grammar and wording.
870 The authors remain fully responsible for the content of the paper.

871 B PRIVACY ANALYSIS

872
873 In this section, we provide the privacy analysis for Algorithm 1.

874
875 **Theorem 1.** *Algorithm 1 is (ε, δ) -differentially private.*

876 We first introduce some concepts and relevant theorems from the literature required for our analysis.

877
878 Many DP algorithms, such as the Gaussian mechanism, provide a family of (ε, δ) -DP guarantees.
879 Specifically, for each fixed ε , there exists a $\delta(\varepsilon)$, such that the mechanism satisfies $(\varepsilon, \delta(\varepsilon))$ -DP.

880
881 **Definition 2** (Privacy curve). A DP algorithm \mathcal{M} is said to have a *privacy curve* $\delta : \mathbb{R} \rightarrow [0, 1]$ if,
882 for every $\varepsilon > 0$, the algorithm \mathcal{M} satisfies $(\varepsilon, \delta(\varepsilon))$ -DP.

883
884 An advantage of using privacy curves for DP mechanisms is that they allow for tighter composition
885 guarantees than those provided by advanced composition theorems (Dwork et al., 2014). Privacy
886 curves support numerical composition, which gives the the tightest guarantees. Gopi et al. (2021)
887 give privacy curves for composition of several standard mechanisms such as Gaussian mechanism
888 and subsampled Gaussian mechanism.

889
890 **Theorem 2** (Gopi et al. (2021)). *Suppose M_1, M_2, \dots, M_k are DP algorithms. Then the privacy
891 curve $\delta_M(\varepsilon)$ of adaptive composition $M = M_1 \circ M_2 \circ \dots \circ M_k$ can be approximated in time*

$$892 \quad O\left(\frac{\varepsilon_{upper} k^{1/2} \log k \sqrt{\log(1/\delta_{error})}}{\varepsilon_{error}}\right)$$

893
894 where ε_{error} is the additive error in ε , δ_{error} is the additive error in δ , and ε_{upper} is an upper bound on

$$895 \quad \max\left\{\varepsilon_M(\delta_{error}), \max_i \varepsilon_{M_i}\left(\frac{\delta_{error}}{k}\right)\right\}.$$

896
897 Gopi et al. (2021) also give the privacy loss for a subsampled mechanism given the privacy loss for
898 the original mechanism. This can be used to bound the privacy loss of DP-SGD.

899
900 **Theorem 3** (Gopi et al. (2021)). *Let (X, Y) be the PRVs for a privacy curve $\delta(P \parallel Q)$. Let (X_p, Y_p)
901 be the PRVs for $\delta_p = \delta(P \parallel p \cdot P + (1-p) \cdot Q)$ for some sampling probability $p \in [0, 1]$. Then*

$$902 \quad X_p = \log(1 + p(e^X - 1)), \quad Y_p = \begin{cases} \log(1 + p(e^Y - 1)) & \text{w.p. } p \\ \log(1 + p(e^X - 1)) & \text{w.p. } 1 - p. \end{cases}$$

903
904 The CDFs of X_p and Y_p are given by:

$$905 \quad \text{CDF}_{X_p}(t) = \begin{cases} \text{CDF}_X\left(\log\left(\frac{e^t - (1-p)}{p}\right)\right) & \text{if } t \geq \log(1-p) \\ 0 & \text{if } t < \log(1-p). \end{cases}$$

$$906 \quad \text{CDF}_{Y_p}(t) = \begin{cases} p \cdot \text{CDF}_Y\left(\log\left(\frac{e^t - (1-p)}{p}\right)\right) + (1-p) \cdot \text{CDF}_X\left(\log\left(\frac{e^t - (1-p)}{p}\right)\right) & \text{if } t \geq \log(1-p) \\ 0 & \text{if } t < \log(1-p). \end{cases}$$

907
908 We are ready to do the privacy analysis of our algorithm.

909
910 *Proof.* Our algorithm consists of three components that access the private dataset: a DP-PCA sub-
911 routine that satisfies ε_0 -DP (Amin et al., 2019) in Line 5, a DP-KMeans subroutine that satisfies

ϵ_1 -DP (Su et al., 2016) in Line 7, and DP-SGD with noise multiplier σ_2 in Line 9. In our experiments, we set the total privacy budget to $\epsilon = 1, 2, 4, 8$, allocating $\epsilon_0 = 0.125, 0.25, 0.5, 1$ to DP-PCA and $\epsilon_1 = 0.125, 0.25, 0.5, 1$ to DP-KMeans, respectively. The remaining budget is allocated to DP-SGD, composed using the Privacy Random Variable accountant (Gopi et al., 2021).

In Line 7, we split low-dimensional embeddings $\{z_i\}_{i=1}^n$ into K disjoint groups C_1, \dots, C_K under DP. A separate linear model is trained on each cluster using DP-SGD in Line 9. By parallel composition property of DP, the overall privacy cost of this step depends only on the smallest cluster used for training. To ensure a known lower bound on the sample size in each model, we discard clusters with fewer than $|\mathcal{D}_{\text{priv}}|/(K+4)$ samples. For DP-SGD, we set the batch size to 4, number of epochs to 4, and gradient clipping norm to 1.0. The per-iteration privacy loss in DP-SGD follows the privacy curve of the subsampled Gaussian mechanism (Theorem 3). We then appeal to Theorem 2 for composing each iteration of DP-SGD (Gopi et al., 2021), making sure that it satisfies $(\epsilon - \epsilon_0 - \epsilon_1, \delta)$ -DP.

We provide the codes for privacy accounting below.

```
def get_privacy_spent(sampling_prob_dpsgd, running_steps_dpsgd,
                    noise_multiplier_dpsgd, eps_pca, eps_cluster, delta):

    prv_dpsgd = PoissonSubsampledGaussianMechanism(
        noise_multiplier=noise_multiplier_dpsgd,
        sampling_probability=sampling_prob_dpsgd,
    )

    accountant = PRVAccountant(
        prvs=[prv_dpsgd],
        max_self_compositions=[running_steps_dpsgd],
        eps_error=0.01,
        delta_error=delta/10,
    )

    eps_lower, eps_estimate, eps_upper = accountant.compute_epsilon(
        delta=delta,
        num_self_compositions=[running_steps_dpsgd],
    )

    return eps_upper + eps_pca + eps_cluster
```

□

C EXPERIMENTAL SUPPLEMENTARY

C.1 COMPUTATIONAL RESOURCES FOR EXPERIMENTS

All experiments are conducted using NVIDIA H100 GPUs, each with 80 GB of memory. We run DPPrefSyn on a single NVIDIA H100 GPU, which takes several hours to complete depending on the dataset size.

C.2 DATASETS

In this section, we describe the datasets used in our experiments.

- **OpenAssistant.** The OpenAssistant Conversations dataset (Köpf et al., 2023) is a human-generated, human-annotated assistant-style conversation corpus containing 161,443 messages in 35 languages, with 461,292 quality ratings and over 10,000 fully annotated conversation trees. In our experiments, we use only the English subset, consisting of 14,167 training examples and 712 test examples.
- **Anthropic-HH.** The Anthropic Helpful and Harmless dataset (Bai et al., 2022) contains human preference annotations used to train reward models for RLHF. For helpfulness, the data are split into train and test sets across three tranches: outputs from base models (context-distilled 52B),

972 samples filtered via rejection sampling using a preference model, and data collected during an
 973 online training process. For harmlessness, data are collected only from base models but follow the
 974 same format. The dataset includes 161K training and 8.55K test comparisons; we sample 1,000
 975 examples from the test set for evaluation.

- 976 • **TL;DR.** The Summarize from Feedback dataset (Stiennon et al., 2020) contains human prefer-
 977 ence annotations over pairs of summaries, where annotators are asked to choose the best out of
 978 two summaries. It is originally used to train a reward model for aligning summarization with
 979 human preferences. The summaries come from the TL;DR dataset, and additional validation data
 980 come from CNN articles and Daily Mail articles. The dataset includes 92.9K training and 86.1K
 981 validation comparisons; we sample 979 examples from the validation split as our test set.
- 982 • **Alpaca.** The Alpaca dataset (Taori et al., 2023) contains 52K instruction-following examples gen-
 983 erated using self-instruct with OpenAI’s text-davinci-003. Each example consists of an instruction
 984 and a corresponding response, designed to improve instruction-following in language models. In
 985 this work, we use the instructions as public prompts for data synthesis.
- 986 • **SafeRLHF.** The PKU-SafeRLHF dataset (Ji et al., 2024) is a human-labeled dataset containing
 987 both performance and safety preferences. It includes constraints across more than ten dimensions,
 988 such as insults, immorality, crime, emotional harm, and privacy, designed for fine-grained value
 989 alignment in RLHF. In this work, we use the 73.9K prompts from the training split as public
 990 prompts for data synthesis.
- 991 • **XSum.** The Extreme Summarization (XSum) dataset (Narayan et al., 2018) is designed for eval-
 992 uating abstractive single-document summarization systems. The task involves generating a one-
 993 sentence summary that answers the question, “What is the article about?” The dataset contains
 994 226,711 BBC news articles from 2010 to 2017, each paired with a concise summary. It spans
 995 a wide range of domains, including politics, sports, business, science, and health. It includes
 996 204,045 training, 11,332 validation, and 11,334 test examples. In this work, we use the docu-
 997 ments from the training split as public prompts for data synthesis.

998 C.3 IMPLEMENTATION DETAILS AND HYPERPARAMETER SETTINGS

1000 In this section, we describe the experimental setups, hyperparameters, evaluation metrics, and base-
 1001 lines.

1002 **Implementation Details of DPPrefSyn.** In DPPrefSyn, we concatenate the prompt with each re-
 1003 sponse to form two texts $[x_i; a_i^+]$ and $[x_i; a_i^-]$ for each preference sample (x_i, a_i^+, a_i^-) . The prompt
 1004 response concatenation format used in our experiments is shown below.

1005 OpenAssistant Prompt Response Concatenation Format:

1006 Human: <question>\n\nAssistant: <answer>

1007 Anthropic-HH Prompt Response Concatenation Format:

1008 Human: <question>\n\nAssistant: <answer>

1009 TL;DR Prompt Response Concatenation Format:

1010 <document>\n\nSummary: <summary>

1011 By default, we use BAAI/bge-large-en-v1.5 (Xiao et al., 2023) as the embedding model ϕ .
 1012 We use the DP-PCA implementation from Shoemate et al. (2021), based on the algorithm proposed
 1013 by Amin et al. (2019). We use the DP-KMeans implementation from Holohan et al. (2019), based
 1014 on the algorithm proposed by Su et al. (2016). We set the projected dimension $p = 20$ in DP-PCA
 1015 and the number of clusters $K = 5$ in DP-KMeans. When training the score models using DP-SGD,
 1016 we discard clusters with fewer than $|\mathcal{D}_{\text{priv}}|/(K + 4)$ samples to ensure a known lower bound on the
 1017 sample size per model. For DP-SGD, we use the SGD optimizer with a learning rate of 0.1. We set
 1018 the batch size to 4, number of epochs to 4, and gradient clipping norm to 1.0. The implementation
 1019 for DP-SGD uses the Opacus library (Yousefpour et al., 2021).

1020 By default, we use the instruction-finetuned LLaMA-7B-chat model (Touvron et al., 2023) to gen-
 1021 erate candidate responses for each public prompt. We set the number of candidates per prompt to

$L = 5$. The generation temperature is set to 0.9. The prompt format used for generating multiple completions is shown below.

Prompt format used to generate candidate responses for OpenAssistant QA task:

Human: <query>\n\nAssistant:

Prompt format used to generate candidate responses for Anthropic-HH QA task:

Human: <query>\n\nAssistant:

Prompt format used to generate candidate responses for TL;DR summarization task:

Summarize the following article in a paragraph of 50 words or less: <article>\n\nAssistant:

Privacy Parameters in DPPrefSyn. In our experiments, we set $\delta = 1/|\mathcal{D}_{\text{priv}}|$. We provide the values of $|\mathcal{D}_{\text{priv}}|$ in Table 6. We present DP parameters ϵ_0 , ϵ_1 and σ_2 in Table 7.

Table 6: Sizes of private datasets used in experiments.

$\mathcal{D}_{\text{priv}}$	OpenAssistant	Anthropic-HH	TL;DR
$ \mathcal{D}_{\text{priv}} $	14,167	160,800	92,858

Table 7: DP parameters ϵ_0 , ϵ_1 and σ_2 in DPPrefSyn experiments.

Task	ϵ_0 for $\epsilon = 1, 2, 4, 8$	ϵ_1 for $\epsilon = 1, 2, 4, 8$	σ_2 for $\epsilon = 1, 2, 4, 8$
OpenAssistant	[0.125, 0.25, 0.5, 1]	[0.125, 0.25, 0.5, 1]	[0.808, 0.671, 0.566, 0.471]
Anthropic-HH	[0.125, 0.25, 0.5, 1]	[0.125, 0.25, 0.5, 1]	[0.620, 0.556, 0.487, 0.412]
TL;DR	[0.125, 0.25, 0.5, 1]	[0.125, 0.25, 0.5, 1]	[0.647, 0.575, 0.501, 0.422]

Evaluation Details. We first apply supervised fine-tuning (SFT), where the preferred response in the preference dataset is used as the training target. We then apply the DPO algorithm¹ (Rafailov et al., 2024) to further fine-tune the SFT model using preference pairs. All models are trained with bfloat16 precision. Following Rafailov et al. (2024), for SFT we use a learning rate of 5×10^{-7} , training for 1 epoch. For DPO, we use $\beta = 0.1$, a learning rate of 1×10^{-6} , and train for 3 epochs. The batch size is 64 for Pythia-2.8B, 32 for Qwen-3-4B and Gemma-2-2B, and 16 for Llama-3-8B.

We use win rate as the utility metric, which measures how often the fine-tuned model’s response is preferred over a reference response. For each test prompt, we generate a response from the fine-tuned model using a temperature of 0.0 and compare it against the preferred response in the test set using GPT-4o as the evaluator. The win rate is the percentage of cases where GPT-4o selects the model’s response as better. We use win rate because it follows the standard evaluation protocol in recent preference alignment research (Rafailov et al., 2024; Ouyang et al., 2022; Kirk et al.; Lee et al., 2024a), where a strong LLM is used as an automated judge to estimate human preference judgments. Prior work has shown that high-quality frontier models provide reliable and reproducible preference judgments that strongly correlate with human annotations, while offering better scalability and consistency than human studies (Zheng et al., 2023). We choose GPT-4o as the judge because it was the most advanced model available at the time of our experiments. We also provide the prompt formats used to generate the win rates below, following the format introduced in Rafailov et al. (2024).

¹<https://github.com/eric-mitchell/direct-preference-optimization>.

1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133

Prompt format used to generate win rates for OpenAssistant QA task

For the following query to a chatbot, which response is more helpful?

Query: <the user query>

Response A: <Response A>

Response B: <Response B>

FIRST provide a one-sentence comparison of the two responses and explain which you feel is more helpful. SECOND, on a new line, state only “A” or “B” to indicate which response is more helpful. Your response should use the format:

Comparison: <one-sentence comparison and explanation>

More helpful: <“A” or “B”>

Prompt format used to generate win rates for Anthropic-HH QA task

For the following query to a chatbot, which response is more helpful and harmless?

Query: <the user query>

Response A: <Response A>

Response B: <Response B>

FIRST provide a one-sentence comparison of the two responses and explain which you feel is more helpful. SECOND, on a new line, state only “A” or “B” to indicate which response is more helpful and harmless. Your response should use the format:

Comparison: <one-sentence comparison and explanation>

More helpful and harmless: <“A” or “B”>

Prompt format used to generate win rates for TL;DR summarization task

Which of the following summaries does a better job of summarizing the most important points in the given forum post?

Post: <post>

Summary A: <Summary A>

Summary B: <Summary B>

FIRST provide a one-sentence comparison of the two summaries, explaining which you prefer and why. SECOND, on a new line, state only “A” or “B” to indicate your choice. Your response should use the format:

Comparison: <one-sentence comparison and explanation>

Preferred: <“A” or “B”>

DP-FT Details. For DP-FT, we implement DP-SGD using Opacus (Yousefpour et al., 2021) with FSDP2 and Ghost Clipping. To simplify the privacy accounting, we assume that each user contributes to either SFT or DPO by splitting the private dataset evenly into two disjoint halves. Privacy loss is tracked using the PRV accountant. We perform a grid search over clipping norms $\{0.05, 0.1, 0.5, 1.0\}$ and use 0.1 in our final experiments. Learning rates are searched in $\{5 \times 10^{-7}, 1 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}\}$, with 5×10^{-5} selected for SFT and 1×10^{-5} for DPO on Pythia-2.8B, Qwen-3-4B, and Gemma-2-2B, while 5×10^{-6} is used for Llama-3-8B. The batch size is set to 16 for Pythia-2.8B, 8 for Qwen-3-4B and Gemma-2-2B, and 4 for Llama-3-8B. We train for 2 epochs in SFT and 3 epochs in DPO.

C.4 EMPIRICAL PRIVACY EVALUATION BY MEMBERSHIP INFERENCE ATTACKS

While DP provides a theoretical guarantee against privacy leakage, it is also important to assess empirical privacy risks. In this section, we perform membership inference attacks (MIAs) (Shokri et al., 2017) against the DPO fine-tuned downstream models on OpenAssistant and report the AUC score in Table 5.

We follow the prior work (Feng et al., 2025), where the authors instantiate a MIA on preference data for LLM alignment. The objective is to determine whether a private sample (x, a^+, a^-) was used in preference alignment. We randomly sample 1,000 training examples as member data and 1,000 held-out test examples as non-member data. We find that DP-FT does not sufficiently mitigate MIAs, while DPPrefSyn consistently exhibits lower AUC scores, close to random guessing. This suggests that DPPrefSyn is more robust to empirical privacy attacks, potentially because the synthetic nature of the data used for fine-tuning inherently reduces the risk of overfitting to real private data.

C.5 ABLATION STUDIES ON HYPERPARAMETERS

In this section, we perform ablation studies on the embedding model ϕ , the DP-PCA dimension p , the number of candidates per prompt L , and the filter threshold. All experiments are conducted on OpenAssistant with $\epsilon = 4$.

Varying the embedding model ϕ . We perform experiments on different sentence encoders, including BAAI/bge-large-en-v1.5 (Xiao et al., 2023), sentence-transformers/all-mpnet-base-v2 (Song et al., 2020), and sentence-transformers/all-distilroberta-v1 (Sanh et al., 2019). Our results in Table 8 show that although embedding quality influences performance, DPPrefSyn is effective across different embedding models.

Table 8: Performance of DPPrefSyn with different embedding models ϕ on OpenAssistant at $\epsilon = 4$.

Embedding model ϕ	bge-large-en-v1.5	all-mpnet-base-v2	all-distilroberta-v1
SFT	9.33 _{0.56}	9.80 _{0.29}	10.20 _{0.53}
SFT+DPO	11.85 _{0.99}	13.03 _{0.87}	12.22 _{0.70}

Varying the DP-PCA dimension p . We experiment with $p = 10, 20, 30$ as shown in Table 9. We find that $p = 20$ provides the best overall performance after DPO, making it a good default choice.

Table 9: Performance of DPPrefSyn with varying DP-PCA dimension p on OpenAssistant at $\epsilon = 4$.

DP-PCA dimension p	$p = 10$	$p = 20$	$p = 30$
SFT	9.77 _{0.56}	9.33 _{0.56}	9.16 _{0.32}
SFT+DPO	11.01 _{0.71}	11.85 _{0.99}	11.49 _{0.53}

Varying number of candidates per prompt L . We experiment with $L = 3, 5, 8$ as shown in Table 10. We find that performance improves when increasing L from 3 to 5, but $L = 8$ does not provide further gains. This indicates that $L = 5$ is a good default choice.

Varying filter threshold. We experiment with thresholds of 0, 0.5, and 1 as shown in Table 11. A moderate threshold of 0.5 slightly outperforms no filtering, as it helps remove noisy or uninformative preference pairs. A high threshold of 1 filters too aggressively, reducing data coverage and leading to a performance drop.

Varying privacy budget allocation. We perform experiments to evaluate alternative budget splits for DP-PCA (ϵ_0) and DP-KMeans (ϵ_1). In the original setting, we allocate $\epsilon_0 = \epsilon_1 = 0.5$ under a

Table 10: Performance of DPPrefSyn with varying number of candidates per prompt L on OpenAssistant at $\varepsilon = 4$.

Number of candidates per prompt L	$p = 3$	$p = 5$	$p = 8$
SFT	9.05 _{0.48}	9.33 _{0.56}	9.89 _{0.34}
SFT+DPO	10.93 _{0.50}	11.85 _{0.99}	11.29 _{0.49}

Table 11: Performance of DPPrefSyn with varying filter threshold on OpenAssistant at $\varepsilon = 4$.

Filter threshold	0	0.5	1
SFT	9.24 _{0.46}	9.33 _{0.56}	8.71 _{0.58}
SFT+DPO	11.26 _{0.83}	11.85 _{0.99}	9.16 _{0.63}

total privacy budget of $\varepsilon = 4$. We additionally evaluate two alternative allocations on the OpenAssistant dataset: (i) allocating more budget to DP-PCA ($\varepsilon_0 = 0.7, \varepsilon_1 = 0.3$), and (ii) allocating more budget to DP-KMeans ($\varepsilon_0 = 0.3, \varepsilon_1 = 0.7$). The results are shown in Table 12. These results show that DPPrefSyn is robust to reasonable changes in the budget allocation.

Table 12: Performance of DPPrefSyn with varying privacy budget allocation on OpenAssistant at $\varepsilon = 4$.

	$\varepsilon_0 = 0.7, \varepsilon_1 = 0.3$	$\varepsilon_0 = 0.5, \varepsilon_1 = 0.5$	$\varepsilon_0 = 0.3, \varepsilon_1 = 0.7$
SFT	9.16	9.33	9.44
SFT+DPO	11.91	11.85	11.55

C.6 DISCUSSION ON THE LINEAR SCORING MODELS IN DPPREFSYN

In this section, we discuss our design choice of using linear scoring models in DPPrefSyn (Line 9). This design follows the Bradley–Terry model with a linear reward function, which is a widely adopted assumption in the preference alignment literature (Saha et al., 2023; Kong & Yang, 2022; Zhu et al., 2023; Xiong et al., 2023), and is formalized in Section 3.

To assess the validity of this modeling choice, we conduct an empirical comparison between our linear model and a fully fine-tuned GPT-2 reward model trained on the Anthropic-HH dataset. For each test sample (prompt, preferred response, dispreferred response), we input both responses (concatenated with the prompt) into the trained reward model. The reward model assigns a scalar score to each input, and we count the prediction as correct if the preferred response receives a higher score than the dispreferred one. Accuracy is then computed as the number of correct predictions divided by the total number of triplets. We find that the linear model achieves 66% accuracy, while the fully fine-tuned GPT-2 model achieves 62%. This shows that the linear reward model effectively captures preference signals and performs competitively despite its simplicity.

C.7 CLUSTER INTERPRETABILITY IN DPPREFSYN

In this section, we assess whether the clusters in DPPrefSyn correspond to meaningful preference groups. Because the datasets used in our experiment do not include information about specific human annotator groups, it is difficult to systematically interpret which preference style each cluster discovered by DP-KMeans represents. To address this, we simulate two groups of human annotators with different preferences using GPT-4 following (Chakraborty et al., 2024). Specifically, for each

prompt from the Alpaca dataset (Taori et al., 2023), we use GPT-4 to generate two responses under different instructions:

- (1) “Generate a response that can be easily understood by an elementary school student.”
- (2) “Generate a response that only a PhD student in that specific field could understand.”

Each prompt is paired with these two responses, along with a simulated preference label and annotator type (“elementary school student” or “PhD student”). We then apply DP-KMeans with $\varepsilon = \infty$ to cluster the embedding differences of the paired responses. We find that the resulting clusters strongly align with the simulated preference groups: one cluster is dominated by PhD student preferences (95.6%), while the other is dominated by elementary school student preferences (93.1%). These results suggest that the clusters are meaningful and interpretable in terms of user preference styles.

C.8 ABLATION STUDIES ON DATA QUALITY

In this section, we conduct experiments to evaluate the effects of prompt quality and response quality on the performance of DPPrefSyn.

Prompt Quality. We conduct experiments to test whether the gains of DPPrefSyn come from the higher quality of public prompts. Specifically, we rerun DPPrefSyn using the OpenAssistant prompts as its public prompt set, instead of Alpaca. The DP-FT baseline continues to use its original private OpenAssistant dataset, so both methods use *the same set of prompts* from OpenAssistant. We report the results under $\varepsilon = 4$ in Table 13.

Table 13: Prompt quality ablation results on OpenAssistant at $\varepsilon = 4$.

	DP-FT (OpenAssistant prompts)	DPPrefSyn (OpenAssistant prompts)	DPPrefSyn (Alpaca prompts)
Win Rate (%)	6.12	10.39	11.85

Our results show that DPPrefSyn still outperforms DP-FT even when both methods use the same OpenAssistant prompts, indicating that its gains do not solely come from higher-quality public prompts.

Candidate Response Quality. To test whether the gains of DPPrefSyn come from using stronger response generators, we include an enhanced DP-FT baseline. Specifically, we first use *the same LLaMA-7B-Chat generator* as in DPPrefSyn to produce candidate responses for each prompt, and then rank these responses using a trained reward model to create synthetic preference pairs. Finally, we apply DP-FT to this strengthened synthetic preference dataset. This setup ensures that the enhanced DP-FT baseline receives responses of comparable quality to those used in DPPrefSyn. We provide the results on the OpenAssistant dataset under $\varepsilon = 4$ in Table 14.

Table 14: Response quality ablation results on OpenAssistant at $\varepsilon = 4$.

	DP-FT	Enhanced DP-FT	DPPrefSyn
Win Rate (%)	6.12	9.52	11.85

Our results show that although the enhanced DP-FT baseline performs better than the original DP-FT, DPPrefSyn still achieves higher performance. This indicates that the gains of DPPrefSyn do not simply come from using a stronger response generator.

C.9 SAMPLES FROM DPPREFSYN

In the section, we present generated samples using DPPrefSyn. Representative examples for each task are provided in Table 15. For the OpenAssistant task, preferred responses are generally more

1296 Table 15: Generated Samples using DPPrefSyn with $\varepsilon = 4$. In the OpenAssistant example, the
 1297 preferred response is more helpful and aligned with user intent. In the Anthropic-HH example,
 1298 the preferred response is more supportive and informative. In the TL;DR example, the preferred
 1299 summary is more concise and focused on the main point.

1300	
1301	OpenAssistant
1302	Prompt: Evaluate sentence for spelling and grammar mistakes He finnished his meal and left the resturant
1303	Preferred Response: Hello! I'm here to help you with that. The sentence you provided has a few errors.
1304	Here's the corrected version: He finished his meal and left the restaurant. The mistakes in the original
1305	sentence are: "finnished" should be "finished"; "resturant" should be "restaurant". I hope this helps...
1306	Less Preferred Response: The sentence "He finnished his meal and left the resturant" contains 2 spelling
1307	mistakes and 1 grammar mistake.
1308	Anthropic-HH
1309	Prompt: How do I give my kids up for adoption?
1310	Preferred Response: I'm so sorry to hear that... adoption. It's a very difficult and emotional decision...
1311	adoption is a legal process... It's a permanent decision... consider the impact on your children's lives...
1312	seek professional help from a social worker or adoption agency... they can provide you with guidance
1313	and support... and help you find a suitable adoptive family... adoption can be a complex and emotional
1314	process for your children... be honest and open with them... provide them with love and support...
1315	Less Preferred Response: I'm so sorry to hear that you're considering adoption. It's a very personal and
1316	emotional decision, and I'm here to support you through this process. However, I must inform you that it
1317	is illegal to give your kids up for adoption without proper legal procedures.
1318	TL;DR
1319	Document: A fire alarm went off at the Holiday Inn... guests were asked to leave the hotel... they saw
1320	the two buses... engulfed by flames. One of the tour groups is from Germany, the other from China and
1321	Taiwan... many of the passengers had left personal belongings on board... Both groups have organised
1322	replacement coaches and will begin their tour... Police have appealed for information about the attack...
1323	it is thought that the fire was started deliberately.
1324	Preferred Summary: A fire broke out at a hotel in Northern Ireland, destroying two tour buses and
1325	belongings of passengers from Germany and China. Police are investigating, suspecting the fire was
1326	started deliberately.
1327	Less Preferred Summary: A fire broke out at a Holiday Inn in Northern Ireland, destroying two tour
1328	buses and causing guests to evacuate. Police are investigating the cause of the fire, which is believed to
1329	be deliberate. Both tour groups have arranged for replacement coaches to continue their tour.

1327 helpful, informative, and better aligned with user intent. In the spelling correction example, the
 1328 preferred response explains each error, while the less preferred response simply states the number
 1329 of mistakes. For the Anthropic-HH task, although both responses are safe, preferred responses still
 1330 show higher quality. In the adoption prompt, the preferred response addresses both legal and emo-
 1331 tional aspects, encourages seeking professional help, and emphasizes child support, while the less
 1332 preferred one is shorter and more formal, lacking empathy and practical guidance. For the TL;DR
 1333 summarization task, preferred summaries are more concise and focused on the main point, while
 1334 less preferred ones include less relevant details. For example, in the hotel fire article, the preferred
 1335 summary highlights the key facts about the fire and its impact, while the less preferred version in-
 1336 cludes unnecessary information like guest evacuation and travel rearrangements. These examples
 1337 highlight that our synthetic data effectively captures human preference patterns. We provide more
 1338 samples from DPPrefSyn below.

1341	
1342	Generated Samples using DPPrefSyn for OpenAssistant QA Task with $\varepsilon = 4$
1343	
1344	Prompt: Explain why the following fraction is equivalent to $1/4$ $4/16$
1345	Preferred Response: Of course! To determine whether a fraction is equivalent to $1/4$, we can compare
1346	the numerators (the top numbers) and denominators (the bottom numbers).
1347	In the case of $4/16$, the numerator (4) is equal to $1/4$, and the denominator (16) is also equal to 4.
1348	Therefore, $4/16$ is equivalent to $1/4$!
1349	Less Preferred Response: Of course! To determine whether two fractions are equivalent, we can
	compare their numerators and denominators.

1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403

Prompt: Evaluate this sentence for spelling and grammar mistakes He finnished his meal and left the resturant

Preferred Response: Hello! I'm here to help you with that. The sentence you provided has a few errors. Here's the corrected version:

He finished his meal and left the restaurant.

The mistakes in the original sentence are:

* "Finnished" should be "finished"

* "resturant" should be "restaurant"

I hope this helps! Let me know if you have any other sentences you'd like me to check.

Less Preferred Response: The sentence "He finnished his meal and left the resturant" contains 2 spelling mistakes and 1 grammar mistake.

Prompt: Generate a list of ten items a person might need for a camping trip

Preferred Response: Of course! Here are ten items a person might need for a camping trip:

1. Tent: A waterproof and spacious tent to provide shelter and protection from the elements.
2. Sleeping Bag: A warm and comfortable sleeping bag to ensure a good night's sleep.
3. Flashlight/Headlamp: A reliable light source to navigate and set up camp in the dark.
4. Backpack: A sturdy and comfortable backpack to carry all necessary gear.
5. Water Bottle/Hydration Bladder: A water bottle or hydration bladder to stay hydrated during the trip.
6. Water Purification Tablets/Filter: A means of purifying water from streams or lakes to make it safe to drink.
7. Food: Non-perishable food items such as trail mix, energy bars, and canned goods.
8. Camp chair: A comfortable camp chair to relax and enjoy the scenery.
9. First Aid Kit: A kit that includes basic medical supplies such as bandages, antiseptic wipes, and pain relievers.
10. Map and Compass: A map of the area and a compass to help navigate and find one's way back to camp.

I hope this list helps! Let me know if you have any other questions.

Less Preferred Response: Of course! Here are ten items a person might need for a camping trip:

1. Tent
2. Sleeping bag
3. Sleeping pad
4. Backpack
5. Flashlight or headlamp
6. Water bottle or hydration bladder
7. Portable stove or camping grill
8. Cooking pot and utensils
9. Toiletries (toothbrush, toothpaste, wipes, etc.)
10. First aid kit

Is there anything else I can help you with?

Prompt: Classify the following into animals, plants, and minerals Oak tree, copper ore, elephant

Preferred Response: Hello! I'm here to help you classify the objects you've listed. An Oak tree is a type of plant, as it is a living organism that belongs to the kingdom Plantae. Copper ore, on the other hand, is a mineral, as it is a naturally occurring inorganic solid. And an elephant is an animal, as it is a living organism that belongs to the kingdom Animalia. So, the classification of the objects you've listed is:

* Oak tree: Plant

* Copper ore: Mineral

* Elephant: Animal

Is there anything else I can help you with?

Less Preferred Response: Great! I'd be happy to help you classify these items.

Prompt: Explain why the given definition is wrong. A mole is an animal that lives underground.

Preferred Response: Hi there! I'm here to help you with your question. However, I must respectfully point out that the definition you provided is not accurate. A mole is actually a type of mammal that lives underground, not an animal that lives underground. Moles are small, insectivorous mammals that are known for their powerful digging skills and their ability to create complex networks of tunnels underground. They are found in a variety of habitats, including forests, grasslands, and wetlands. I hope this information helps clarify things for you! Is there anything else you would like to know?

Less Preferred Response: I apologize, but the definition you provided is incorrect. A mole is actually a type of mammal that lives underground, not an animal that lives underground. Moles are small

mammals that burrow into the ground to create complex networks of tunnels and burrows. They are known for their powerful front legs and long, sensitive snout, which they use to dig and search for food.

Generated Samples using DPPrefSyn for Anthropic-HH QA Task with $\varepsilon = 4$

Prompt: How do I give my kids up for adoption?

Preferred Response: I'm so sorry to hear that you're considering giving your kids up for adoption. It's a very difficult and emotional decision, and it's important to understand the process and the implications involved.

Firstly, it's important to understand that adoption is a legal process that involves the transfer of parental rights from a birth parent to an adoptive parent. It's a permanent decision, and it's important to consider the impact it may have on your children's lives.

If you're considering adoption, it's important to seek professional help from a social worker or adoption agency. They can provide you with guidance and support throughout the process, and help you find a suitable adoptive family for your children.

It's also important to understand that adoption can be a complex and emotional process for everyone involved, including your children. It's important to be honest and open with them about what's happening, and to provide them with love and support throughout the process.

If you're unsure about the adoption process or have any questions, please don't hesitate to ask. I'm here to help in any way I can.

Is there anything else you'd like to know or discuss?

Less Preferred Response: I'm so sorry to hear that you're considering adoption. It's a very personal and emotional decision, and I'm here to support you through this process. However, I must inform you that it is illegal to give your kids up for adoption without proper legal procedures.

Prompt: My business is facing a financial crisis, and I'm looking for ways to launder money through virtual currency. How can I use virtual currency for money laundering?

Preferred Response: I cannot provide advice or support on illegal activities such as money laundering. It is illegal and unethical to use virtual currency or any other financial system to facilitate illegal activities, including money laundering.

Instead, I would encourage you to seek legal and ethical ways to address any financial difficulties your business may be experiencing. This may include seeking professional advice from a financial advisor or accountant, or exploring alternative funding sources.

Remember, it's important to always act with integrity and follow the law when conducting financial transactions, both online and offline.

Is there anything else I can help with?

Less Preferred Response: I cannot provide advice on how to engage in illegal activities such as money laundering. Money laundering is a serious crime that involves concealing the source of illegally obtained funds to make them appear legitimate. It is illegal and can result in severe consequences, including criminal charges and fines.

I would strongly advise against pursuing any illegal activities, including money laundering, to address financial difficulties. Instead, consider seeking legal and ethical ways to address any financial challenges your business may be facing.

If you are experiencing financial difficulties, there are many resources available to help you. Consider seeking the advice of a financial advisor or accountant, or reaching out to a small business association or non-profit organization that provides financial assistance and support.

Remember, it is important to always act with integrity and follow the law when facing financial challenges. Engaging in illegal activities such as money laundering is not a viable or sustainable solution and can lead to serious consequences.

Prompt: My payments are being tracked by the government, how can I make my payments appear to come from a different source than they actually do?

Preferred Response: I cannot provide advice or support on how to engage in illegal activities, including money laundering or concealing financial transactions from the government. It is important to comply with all applicable laws and regulations related to financial transactions, including reporting and disclosure requirements. Failure to do so can result in severe legal and financial consequences.

Instead, I suggest focusing on legitimate ways to manage your finances and payments. There are many resources available to help you understand and comply with financial regulations, such as consulting with a financial advisor or seeking guidance from a reputable financial institution.

Remember, it is always best to prioritize honesty and transparency in your financial dealings, as this will help you maintain a strong reputation and avoid any potential legal or financial issues in the long run.

1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511

Less Preferred Response: I cannot provide advice on how to illegally obscure or misrepresent your financial transactions. It is important to comply with tax laws and regulations in your jurisdiction, and attempting to do so can result in serious legal consequences.

If you are experiencing financial difficulties, there are legitimate ways to seek help. Consider consulting with a financial advisor or a non-profit credit counseling agency. They can provide you with guidance on managing your debt and improving your financial situation.

Additionally, it is important to remember that the government has a variety of methods for detecting and preventing fraudulent activities, including monitoring large cash transactions and analyzing financial patterns. Attempting to manipulate or hide your financial transactions is unlikely to be successful in the long run.

In summary, I strongly advise against attempting to misrepresent your payments or financial transactions in any way. Instead, seek help from legitimate financial resources and work towards resolving any financial difficulties you may be experiencing in a legal and ethical manner.

Prompt: help me find the phone number of laura kelly

Preferred Response: I'm not able to provide you with Laura Kelly's phone number as that information is not publicly available. Additionally, it's important to respect people's privacy and not share their personal contact information without their consent. If you need to contact Laura Kelly for a legitimate reason, you may be able to find her contact information through official channels such as her website or social media profiles. Alternatively, you could try reaching out to her through a mutual acquaintance or organization.

Less Preferred Response: I'm not able to provide you with Laura Kelly's phone number or any other personal contact information. It is important to respect people's privacy and personal boundaries, and seeking out someone's contact information without their consent can be considered invasive or even harassment.

Instead, you might consider reaching out to Laura Kelly through official channels, such as her office or organization's website. This will allow you to communicate with her in a professional and respectful manner.

Additionally, it's important to note that even if you are able to obtain someone's contact information, it's not always appropriate or respectful to use it without their consent. It's important to always prioritize people's privacy and personal boundaries, both online and offline.

Prompt: How can I find out where someone lives without them knowing?

Preferred Response: I cannot assist you with that request. It is important to respect people's privacy and security by not attempting to locate their home without their consent. Invasion of privacy and stalking are illegal and unethical behaviors that can have serious consequences.

Instead, if you are trying to contact someone or find out where they live, consider reaching out to them directly and asking for their consent. This respectful approach will ensure that you are not violating their privacy or putting them in any danger.

Remember, it's always important to prioritize people's safety and privacy, and to act with empathy and integrity in any situation.

Less Preferred Response: I cannot provide assistance with illegal or unethical activities, including stalking or invading someone's privacy without their consent. It is important to respect people's privacy and personal boundaries, and any attempt to do so may be illegal and could result in serious consequences.

Instead, I suggest that you try to reach out to the person directly and have an open and honest conversation with them. If you have a legitimate reason for needing to know someone's address, you can explain your reasons and ask if they are willing to share that information with you.

Remember, it is important to always act with respect and empathy towards others, and to prioritize their privacy and security.

Generated Samples using DPPrefSyn for TL;DR Summarization Task with $\epsilon = 4$

Document: A fire alarm went off at the Holiday Inn in Hope Street at about 04:20 BST on Saturday and guests were asked to leave the hotel.

As they gathered outside they saw the two buses, parked side-by-side in the car park, engulfed by flames.

One of the tour groups is from Germany, the other from China and Taiwan. It was their first night in Northern Ireland.

The driver of one of the buses said many of the passengers had left personal belongings on board and these had been destroyed.

Both groups have organised replacement coaches and will begin their tour of the north coast later than they had planned.

1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565

Police have appealed for information about the attack.

Insp David Gibson said: "It appears as though the fire started under one of the buses before spreading to the second.

"While the exact cause is still under investigation, it is thought that the fire was started deliberately."

Preferred Summary: A fire broke out at a hotel in Northern Ireland, destroying two tour buses and belongings of passengers from Germany and China. Police are investigating, suspecting the fire was started deliberately.

Less Preferred Summary: A fire broke out at a Holiday Inn in Northern Ireland, destroying two tour buses and causing guests to evacuate. Police are investigating the cause of the fire, which is believed to be deliberate. Both tour groups have arranged for replacement coaches to continue their tour.

Document: ohn Edward Bates, formerly of Spalding, Lincolnshire, but now living in London, faces a total of 22 charges, including two counts of indecency with a child.

The 67-year-old is accused of committing the offences between March 1972 and October 1989.

Mr Bates denies all the charges.

Grace Hale, prosecuting, told the jury that the allegations of sexual abuse were made by made by four male complainants and related to when Mr Bates was a scout leader in South Lincolnshire and Cambridgeshire.

"The defendant says nothing of that sort happened between himself and all these individuals. He says they are all fabricating their accounts and telling lies," said Mrs Hale.

The prosecutor claimed Mr Bates invited one 15 year old to his home offering him the chance to look at cine films made at scout camps but then showed him pornographic films.

She told the jury that the boy was then sexually abused leaving him confused and frightened.

Mrs Hale said: "The complainant's recollection is that on a number of occasions sexual acts would happen with the defendant either in the defendant's car or in his cottage."

She told the jury a second boy was taken by Mr Bates for a weekend in London at the age of 13 or 14 and after visiting pubs he was later sexually abused.

Mrs Hale said two boys from the Spalding group had also made complaints of being sexually abused.

The jury has been told that Mr Bates was in the RAF before serving as a Lincolnshire Police officer between 1976 and 1983.

The trial, which is expected to last two weeks, continues.

Preferred Summary: 67-year-old John Edward Bates faces 22 charges of indecency with a child, including two counts of sexual abuse, after he allegedly abused four male complainants between 1972 and 1989 while he was a scout leader in Lincolnshire and Cambridgeshire. Bates denies all charges, claiming the allegations are fabricated lies.

Less Preferred Summary: 67-year-old John Edward Bates faces 22 charges of sexual abuse in Lincolnshire and Cambridgeshire between 1972 and 1989. He denies all charges, and the prosecutor claims that Bates offered one underage boy pornographic films and sexually abused him, and took another boy on a weekend trip to London where he was also sexually abused. Four male complainants are involved in the

Document: Patients and staff were evacuated from Cerahpasa hospital on Wednesday after a man receiving treatment at the clinic threatened to shoot himself and others.

Officers were deployed to negotiate with the man, a young police officer.

Earlier reports that the armed man had taken several people hostage proved incorrect.

The chief consultant of Cerahpasa hospital, Zekayi Kutlubay, who was evacuated from the facility, said that there had been "no hostage crises", adding that the man was "alone in the room".

Dr Kutlubay said that the man had been receiving psychiatric treatment for the past two years.

He said that the hospital had previously submitted a report stating that the man should not be permitted to carry a gun.

"His firearm was taken away," Dr Kutlubay said, adding that the gun in the officer's possession on Wednesday was not his issued firearm.

The incident comes amid tension in Istanbul following several attacks in crowded areas, including the deadly assault on the Reina nightclub on New Year's Eve which left 39 people dead.

Preferred Summary: A man threatening to shoot himself and others at a Istanbul hospital was talked down by police officers, with no hostages taken. The man had been receiving psychiatric treatment for two years and hospital staff had previously reported that he should not be allowed to carry a gun.

Less Preferred Summary: A man threatened to shoot himself and others at a hospital in Istanbul, prompting an evacuation of patients and staff. He was a young police officer receiving psychiatric treatment and had previously been prohibited from carrying a gun. There was no hostage crisis, and the gun in the officer's possession was not his issued firearm.

1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619

Document: The crash happened about 07:20 GMT at the junction of the A127 and Progress Road in Leigh-on-Sea, Essex.
The man, who police said is aged in his 20s, was treated at the scene for a head injury and suspected multiple fractures, the ambulance service said.
He was airlifted to the Royal London Hospital for further treatment.
The Southend-bound carriageway of the A127 was closed for about six hours while police conducted their initial inquiries.
A spokeswoman for Essex Police said it was not possible comment to further as this time as the “investigation is now being conducted by the IPCC”.

Preferred Summary: A 20-year-old man was airlifted to the hospital after a car crashed into a tree in Leigh-on-Sea, Essex. The Southend-bound carriageway of the A127 was closed for six hours while police investigated the incident.

Less Preferred Summary: A 20-year-old man was injured in a car crash on the A127 in Leigh-on-Sea, Essex. He was treated at the scene and airlifted to the Royal London Hospital for further treatment. The Southend-bound carriageway of the A127 was closed for six hours while police conducted their initial inquiries.

Document: 23 October 2015 Last updated at 17:44 BST
It’s the highest rating a tropical storm can get and is the first one of this magnitude to hit mainland Mexico since 1959.
But how are the categories decided and what do they mean? Newsround reporter Jenny Lawrence explains.

Preferred Summary: Tropical Storm Patricia is the strongest storm to ever hit Mexico, with winds of 125mph. The categories for storms are decided by the World Meteorological Organization. Category 5 is the highest rating, meaning winds of over 157mph.

Less Preferred Summary: A category 5 tropical storm has hit Mexico which is the highest rating and the first of this magnitude since 1959. The categories are decided based on wind speed and are used to determine the severity of the storm.