Robust Utility-Preserving Text Anonymization Based on Large Language Models

Anonymous ACL submission

Abstract

Text anonymization is crucial for sharing sensitive data while maintaining privacy. Existing techniques face the emerging challenges of re-identification attack ability of Large Language Models (LLMs), which have shown advanced capability in memorizing detailed information and patterns as well as connecting disparate pieces of information. In defending against LLM-based re-identification attacks, anonymization could jeopardize the utility of the resulting anonymized data in downstream tasks—the trade-off between privacy and data utility requires deeper understanding within the context of LLMs. This paper proposes a framework composed of three LLM-based components—a privacy evaluator, a utility evaluator, and an optimization component, which work collaboratively to perform anonymization. To provide a practical model for large-scale and real-time environments, we distill the anonymization capabilities into a lightweight model using Direct Preference Optimization (DPO). Extensive experiments demonstrate that the proposed models outperform baseline models, showing robustness in reducing the risk of re-identification while preserving greater data utility in downstream tasks.[1]

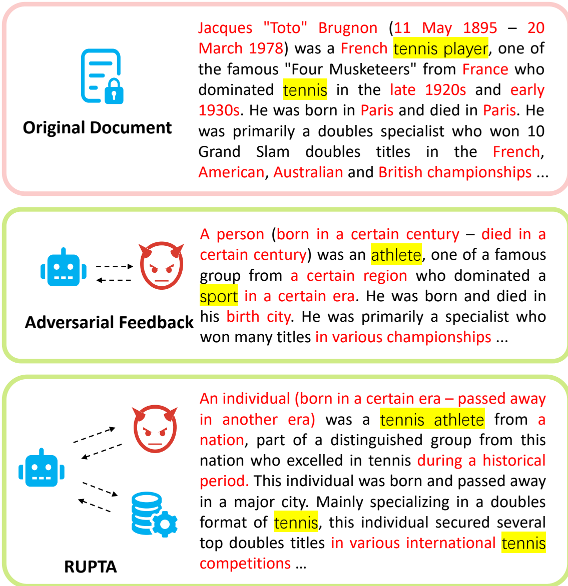


Figure 1: Anonymization examples of the Adversarial Feedback (Staab et al., 2024b) (middle box) and the proposed RUPTA (bottom box) model. The red fonts mark the personally identifiable information. We highlight entities that are critical for our downstream task: occupation classification.

1 Introduction

Privacy protection is a fundamental societal value, enforced through various legal frameworks, e.g., the General Data Protection Regulation (GDPR) in the European Union and the California Consumer Privacy Act (CCPA) in the United States (Voigt and Von dem Bussche, 2017), among others. The recent advancement in large language models (LLMs) and artificial intelligence (AI) tools present both challenges and opportunities in achieving the goal.

Text anonymization is a critical method for safeguarding private and sensitive information. However, current techniques are vulnerable to disclosure threats from increasingly sophisticated Large Language Models (LLMs). Recent studies have demonstrated that these models can re-identify private information, even from texts anonymized by advanced methods (Patsakis and Lykousas, 2023; Staab et al., 2024a; Nyffenegger et al., 2023).

The first key challenge and requirement is, therefore, defending against LLM-based re-identification attacks. In defending against these powerful models, the anonymization process may compromise the utility of the resulting anonymized data in downstream tasks (Mozes and Kleinberg, 2021; Patsakis and Lykousas, 2023). As shown in Fig. 1, while the current state-of-the-art (SoTA) method, which conducts anonymization based on iterative refining according to feedback from a simulated attacker (Staab et al., 2024b), can defend against re-identification attack well, it may eliminate the information crucial for the downstream

[1] Our code and dataset will be released at github.

task. We believe that the trade-off between privacy and data utility requires deeper understanding within the context of LLMs, in which LLMs' re-identification capacity challenges the existing anonymization models, while if properly utilized, LLMs can help build more capable anonymization components to mitigate the discussed adversaries.

In this paper, we introduce a novel framework named Robust Utility-Preserving Text Anonymization (RUPTA), consisting of a *privacy evaluator* (*P-Evaluator*), a *utility evaluator* (*U-Evaluator*), and an *optimization component*. These components are built on LLMs, where the *P-Evaluator* assesses re-identification risks and provides guidance to enhance anonymization robustness against re-identification attacks, the *U-Evaluator* gauges downstream tasks' performance to indicate the level of preserved utility, and the *optimization component* iteratively edits the text based on these evaluation results to jointly optimize both objectives until pre-defined conditions are met. As shown in Fig. 1, RUPTA can ensure privacy-preserving performance comparable to the SOTA method while retaining critical information necessary for accurately classifying the text pertaining to a Tennis Player.

The anonymization models based on LLMs often rely heavily on time-consuming and resource-intensive interactions with LLMs, making these models less feasible for large-scale or real-time applications. To mitigate this problem, we distill the anonymization capabilities into a lightweight model. Our experiments show that the fine-tuned lightweight model achieves a performance comparable to GPT-4, and utilizing Direct Preference Optimization (DPO) (Rafailov et al., 2023) enhances the anonymization efficacy. Our main contributions are summarized as follows:

- To the best of our knowledge, this is the first work to simultaneously optimize privacy and utility in text anonymization using SoTA LLMs, which is crucial for real-life applications.
- We propose a novel framework for text anonymization that is built on the powerful ability of LLMs, consisting of a privacy evaluator, utility evaluator, and optimizer component, which work jointly to perform anonymization and show superior performance over the baseline models.
- We develop more practical methods based on DPO to distill the anonymization capabilities into lightweight models with performance comparable to the teacher models.
- We create a new dataset using the celebrity biographies from DBpedia (Dan, 2019) with occupation labels, serving as a practical benchmark for evaluating the impact of anonymization methods on utility. Anonymization results from LLMs are also included to aid future text anonymization research.

2 Related Work

Text Anonymization. The task is primarily addressed through natural language processing (NLP) and privacy-preserving data publishing (PPDP) approaches. NLP methods use sequence labeling models trained on manually annotated data to identify and remove pre-defined categories of sensitive entities, such as names and phone numbers (Hathurusinghe et al., 2021; Francopoulo and Schaub, 2020). Rather than masking entities according to the pre-defined categories, the PPDP-based approaches mask entities according to the disclosure risk calculated through a privacy model defined by domain experts (Sánchez and Batet, 2016, 2017). However, most existing studies either neglect the utility of anonymized text for downstream tasks or only evaluate it post-anonymization (Yermilov et al., 2023; Staab et al., 2024b), complicating the identification of a strategy that optimally balances privacy and utility. Furthermore, commonly used datasets (Lebret et al., 2016; Pilán et al., 2022) in this field often lack labels for specific downstream tasks, rendering it difficult to assess the impact of anonymization operations on them.

LLMs as the Black-box Optimizer Optimization entails the iterative generation and evaluation of solutions to enhance a specific objective function. Leveraging their robust knowledge storage and generation capabilities, LLMs can identify optimal solutions for intricate real-world optimization problems through effective prompting without necessitating additional training (Prasad et al., 2023; Zhou et al., 2023). In the context of multi-objective optimization problems (MOPs), which involve two or more conflicting objectives, current methodologies typically combine Evolutionary Algorithms with LLMs (Yang and Li, 2023). This approach, however, requires numerous objective evaluations, rendering it impractical for scenarios where evaluating objectives is costly. Our proposed RUPTA serves as an alternative when the preference over objectives is pre-defined.

3 Our Approach

In this section, we present our proposed RUPTA framework, which protects the privacy of the sensitive text while maintaining its utility for analytical purposes. The overview of our framework is shown in Fig. 2. Given a span of text $\boldsymbol{x}_0$, RUPTA iteratively refines the anonymized text to optimize the privacy and utility objectives simultaneously. At iteration $t+1$, the previously anonymized text $\boldsymbol{x}_t$ is input into the system, as shown in the bottom left of Fig. 2. The privacy evaluator (P-Evaluator) analyzes $\boldsymbol{x}_t$ to determine its privacy protection level based on the ground-truth personal information y and provide feedback to enhance its robustness against re-identification attacks. The utility evaluator (U-Evaluator) assesses its usefulness for the downstream tasks based on the corresponding ground-truth label c. Feedback from both evaluators is then used by the optimizer to refine the text using available editing operations, producing the updated text $\boldsymbol{x}_{t+1}$, as shown in the top right of Fig. 2. Specific content of the involved instructions can be found in Appx. C.1.

3.1 Problem Formulation

The text anonymization challenge can be recast as a multi-objective optimization problem with two conflicting objectives: privacy and utility. In this context, privacy should be prioritized over utility. This hierarchy is established by ordering the objectives, transforming the problem into a lexicographic optimization issue (Zykina, 2004). The primary objective is to maximize the level of privacy preservation, ensuring that sensitive information is well-protected against re-identification risks. The secondary objective is to maintain as much useful information as possible in the anonymized text for analytical tasks. This lexicographic optimization problem can be formally expressed as

$$\begin{aligned} &\text{lex max } F(\boldsymbol{x}) = [f_p(\boldsymbol{x}), f_u(\boldsymbol{x})] \\ &\text{St. } \boldsymbol{x} \in \mathcal{X}_0 \end{aligned} \tag{1}$$

where $f_p(\cdot)$ and $f_u(\cdot)$ denote the privacy and utility objective function, respectively. $\mathcal{X}_0$ denotes the set of all possible edits of $\boldsymbol{x}_0$. A solution $\boldsymbol{x}_a \in \mathcal{X}_0$ is lexicographically preferable to another solution $\boldsymbol{x}_b \in \mathcal{X}_0$, denoted as $\boldsymbol{x}_a \succ_{\text{lex}} \boldsymbol{x}_b$, if and only if $f_p(\boldsymbol{x}_a) > f_p(\boldsymbol{x}_b)$ or $(f_p(\boldsymbol{x}_a) = f_p(\boldsymbol{x}_b)$ and $f_u(\boldsymbol{x}_a) > f_u(\boldsymbol{x}_b))$. To solve this lexicographic optimization problem, we propose RUPTA, an iterative method with LLMs to generate, evaluate, and optimize the anonymized text.

3.2 The Privacy Evaluator

Algorithm 1 Privacy Objective Evaluation f_p

Input Anonymized text $\boldsymbol{y}_t$, ground-truth personal information x, instruction $\mathbf{I}_p$, P-Evaluator $\mathcal{LLM}(\cdot)$
Output Privacy objective value p_t and textual feedback $\boldsymbol{f}_t$

1: $(y'_1, y'_2, ..., y'_K) \sim \mathcal{LLM}(\mathbf{I}_p || \boldsymbol{x}_t)$
2: **if** y in $(y'_1, y'_2, ..., y'_K)$ **then**
3: $\quad p_t \leftarrow$ rank of y in $(y'_1, y'_2, ..., y'_K)$
4: $\quad \boldsymbol{f}_t \sim \mathcal{LLM}(\mathbf{I}_{pa} || \boldsymbol{x} || y)$
5: **else**
6: $\quad p_t \leftarrow K+1$
7: $\quad \boldsymbol{f}_t \leftarrow \emptyset$
8: **end if**

The role of the *Privacy Evaluator* (*P-Evaluator*) is to assess the privacy protection level of the anonymized text, ensuring that private content is adequately obscured against re-identification. Besides, it is essential to provide textual feedback to the LLM optimizer as guidance (Pryzant et al., 2023). Thus, the privacy objective evaluation process $f_p(\cdot)$ is formally defined as

$$\boldsymbol{f}_t, p_t = f_p(\boldsymbol{x}_t) \tag{2}$$

where p_t denotes the value of the privacy objective and $\boldsymbol{f}_t$ denotes the textual feedback. We describe the detailed process of privacy evaluation in Alg. 1.

P-Evaluator is instantiated as an LLM. Given the anonymized text $\boldsymbol{x}_t$, we concatenate it with the privacy inference instruction $\mathbf{I}_p$ as input to prompt the P-Evaluator to semantically infer the personal information as shown in *line 1* of Alg. 1, where $||$ denotes concatenation. This step generates top-K inference results $[y'_i]_1^K$ for the personal information. Each result is then compared with the ground-truth personal information y. If a match is found within these top-K results, its rank is used as the scalar privacy score p_t. Further, the evaluator is prompted to provide natural language feedback $\boldsymbol{f}_t$ detailing the clues that led to the correct inference. Otherwise, we set the p_t as $K+1$, representing the maximum achievable score for the privacy objective.

The scalar score p_t quantifies the privacy risk associated with the anonymized text, while the textual feedback $\boldsymbol{f}_t$ offers qualitative insights, guiding the lexicographic optimizer on how to better obscure identifiable information. The value of K serves as a customizable parameter that adjusts the

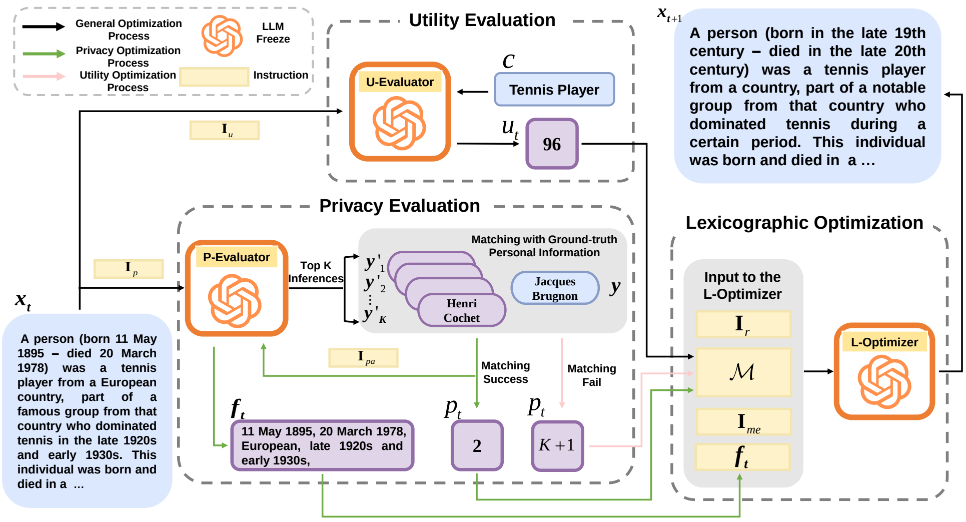


Figure 2: The framework of our proposed RUPTA method

sensitivity of the privacy evaluation, with higher values indicating a more inclusive search for potential privacy breaches, thus facilitating a manually adjustable trade-off between privacy and utility.

3.3 The Utility Evaluator

The *Utility Evaluator* (*U-Evaluator*) is used to ensure that the anonymized text retains its utility for specific analytical tasks, a critical consideration for practical applications across various domains. It analyzes the anonymized text $\boldsymbol{x}_t$, specifically assessing its effectiveness in supporting accurate occupation classification c. The formal utility objective evaluation process is defined as

$$u_t = f_u(\boldsymbol{x}_t, c) \tag{3}$$

where u_t is the utility objective value.

In this paper, we instantiate the U-evaluator with an LLM. Given the anonymized text $\boldsymbol{x}_t$ and the corresponding ground-truth occupation label c, the LLM-based U-evaluator follows the instruction $\mathbf{I}_u$ to output a confidence score u_t

$$u_t \sim \mathcal{LLM}(\mathbf{I}_u || \boldsymbol{x}_t || c), \tag{4}$$

this confidence score quantifies the evaluator's uncertainty that $\boldsymbol{x}_t$ can be correctly classified into the ground truth occupation category c_t, reflecting the degree to which key utility information is preserved.

To better align feedback with real-world use scenarios, the U-Evaluator can be instantiated with the actual model employed in the downstream task. For example, if the anonymized text is intended for sentiment analysis, the U-Evaluator can be instantiated with a sentiment analysis model. The utility score u_t can then be calculated through the logit of the ground-truth label following the traditional uncertainty quantification method (Sensoy et al., 2021).

3.4 Lexicographic Optimizer

Lexicographic optimization (LO) is a special case of MOPs where multiple conflicting objectives are to be maximized simultaneously. In LO, objectives are ranked in order of importance, enabling prioritization of the most critical objectives. The LO problem is generally solved by the sequential optimization method (Zykina, 2004; Zhang et al., 2022). Specifically, regarding the text anonymization problem, privacy and utility are the two objectives, and privacy should be prioritized.

RUPTA employs the LLM as a black-box lexicographic optimizer in a zero-shot manner, where the LLM is prompted to achieve better solutions incrementally based on the history of optimization results and objective evaluations. The overall prompt consists of the pre-defined overall optimization description prompt $\mathbf{I}_r$, the memory module $\mathcal{M}$, the meta instruction variable $\boldsymbol{I}_{me}$ and the textual feedback $\boldsymbol{f}_t$ from P-Evaluator. The memory module $\mathcal{M}$ stores history optimization results and their corresponding privacy and utility objective values. Formally, we have $\mathcal{M} = \{(\boldsymbol{x}_i, p_i, u_i, r_i) | i = 1, 2, ..., t\}$.

To ensure that the primary goal of achieving maximum privacy is prioritized and only after the privacy objective is satisfactorily met does the optimizer focus on improving utility, the lexicographic-optimizer LLM operates in two different modes. When the privacy objective value has not yet reached the pre-set maximum $K+1$, the lexicographic optimizer should focus on maximizing the privacy objective, which is achieved by taking the value of meta instruction variable as $\mathbf{I}_{pr}$ that instructs the LLM to further anonymize $\boldsymbol{x}_t$ according to the textual feedback $\boldsymbol{f}_t$. The process can be formulated as

$$\boldsymbol{x}_{t+1} \sim \mathcal{LLM}(\mathbf{I}_r || \mathcal{M} || \mathbf{I}_{pr} || \boldsymbol{f}_t) \tag{5}$$

Once the privacy objective value has reached the maximum threshold, the meta instruction shifts to $\mathbf{I}_{ur}$, prompting the LLM to optimize the utility level without compromising the achieved privacy objective value.

$$\boldsymbol{x}_{t+1} \sim \mathcal{LLM}(\mathbf{I}_r || \mathcal{M} || \mathbf{I}_{ur}) \tag{6}$$

This iterative process continues until either the pre-defined maximum values for both objectives are reached or the maximum number of iterations T is met. By continuously refining and evaluating the anonymized text, the optimizer iteratively improves it to achieve an optimal balance between privacy and utility.

3.5 Distilling the Anonymization Ability

Utilizing LLMs for text anonymization is computationally expensive, and for certain LLMs, access is only available through APIs, which raises privacy and cost concerns. However, in our framework, the optimization result heavily depends on the reasoning ability of the LLM, which stems from the large scale of parameters these models possess. Recent studies have demonstrated that prompting LLMs as optimizers is less effective with smaller-scale models (Zhang et al., 2024).

To address this issue, we employ knowledge distillation (KD), where a large model (the teacher) transfers its knowledge to a smaller model (the student). Typically, KD involves training the student model using the outputs of the teacher model as labels (Kim and Rush, 2016). In our case, we utilize the final anonymization result produced by the teacher model during the lexicographic optimization as the training label for the student model.

To utilize the generation results of the teacher model more efficiently, we adopt the Direct Preference Optimization (DPO) (Rafailov et al., 2023) method. This method fine-tunes an LLM on human labels of the relative quality of model generations to align the model with human preferences. In our method, intermediate optimization results from the teacher model can be considered less preferred than the final optimization result. These intermediate and final results form the preference dataset. We fine-tune the student model using DPO on this dataset to preferentially generate outputs similar to the final optimization result while reducing the likelihood of producing results akin to the intermediate stages.

Dataset	#Train	#Validation	#Test
DBPedia Classes	1938	243	239
Personal Reddit	318	-	207

Table 1: Statistics of experiment datasets.

4 Experimental Set-up

Datasets. We evaluate our text anonymization method on the following two datasets:

- We sampled celebrity biographies from the DBpedia Classes dataset (Dan, 2019) to build a new dataset **DB-bio**. Unlike the commonly-used Wiki-bio dataset (Lebret et al., 2016) in anonymization studies that lacks annotations for downstream tasks, this dataset includes detailed three-level hierarchical category annotations. We use the third-level category labels as occupation classification labels to assess the impact of our anonymization method on this specific downstream task. The name of the person described by the biography is used as the ground-truth personal information.
- To further validate the generality of our method, we evaluate it on the **PersonalReddit** (PR) dataset (Staab et al., 2024a) consisting of 525 human-verified synthetic public Reddit comments and corresponding user profiles. We use the annotated occupation attribute in the profile as the label of the occupation classification task and anonymize the comments to prevent the identification of other personal attributes.

General statistics of these datasets can be seen in Tab. 1. Detailed statistics, including category distributions, are provided in Appx. A.

Evaluation Metrics. To evaluate our text anonymization method, we focus on two critical aspects: disclosure risk and utility preservation. Disclosure risk is assessed by measuring the **Success Rate** (SR) of a strong adversarial LLM in inferring personal information from anonymized text. Additionally, we prompted an LLM to generate the **Confidence Scores** (CS), evaluating the degree

	Method	Disclosure Risk		Utility Preserving				
		SR$\Downarrow$	CS$\Downarrow$	Precision$\Uparrow$	Recall$\Uparrow$	F1$\Uparrow$	Accuracy$\Uparrow$	Loss$\Downarrow$
DB-bio	**Original**	100.00	98.45	99.58	99.68	99.61	99.58	0.0422
	Azure (Aahill, 2023)	78.24	80.87	91.63	95.04	92.39	92.47	0.3202
	DEID-GPT (Liu et al., 2023)	77.10	79.47	90.82	94.37	92.56	91.22	0.3103
	SD (Dou et al., 2023)	73.21	73.63	92.27	93.11	92.69	92.96	0.2719
	AF (Staab et al., 2024b)	52.91	**50.84**	91.20	94.26	91.75	92.02	0.4048
	RUPTA (Mixtral 8×22b)	67.78	67.15	**96.18**	**97.13**	**96.30**	**96.23**	0.2167
	RUPTA (Llama-3-70b)	64.02	63.23	95.34	96.23	95.55	95.82	0.2224
	RUPTA (GPT-3.5)	68.51	69.16	95.40	96.02	95.70	95.49	0.2188
	RUPTA (GPT-4)	**52.67**	53.11	95.58	96.26	95.91	96.02	**0.1618**

Table 2: Main experiment results on the test set of DB-bio dataset. The top and second performance are highlighted with bold font and underline, respectively.

of confidence with which anonymized text can be linked to the ground-truth personal information.

Utility preservation metrics are gauged by the performance of a simple neural network classifier trained on non-anonymized train data and tested on anonymized text, including **Accuracy**, **Precision**, **Recall**, **F1 Score**, and the classifier's **loss function value** indicating classification uncertainty. Specific metric settings can be seen in Appx. B.

Comparison Methods. To establish the effectiveness of our text anonymization framework, we benchmark it against state-of-the-art methods and industry standards.

- We use **Azure** (Aahill, 2023)'s industry-standard state-of-the-art text anonymizer as a traditional anonymization baseline.
- **AF** (Staab et al., 2024b) is a current state-of-the-art method for text anonymization based on the adversarial feedback mechanism.
- **DEID-GPT** (Liu et al., 2023) prompts the LLM to mask out all the entities of pre-defined kinds.
- **SD** (Dou et al., 2023) prompts the LLM to replace the entities of pre-defined kinds with more general counterparts.

All these methods are recreated using the GPT-4 model (Achiam et al., 2023). Besides, we explore the effectiveness of using different LLM architectures as the lexicographic optimizer, including open-sourced models like instruction-tuned Llama-3-70b (AI@Meta, 2024) and Mixtral 8 × 22b (Jiang et al., 2024), and the proprietary GPT-4 and GPT-3.5. Besides, we evaluate the original non-anonymized dataset (**Original**) for reference.

Implementation Details. GPT-4 is used exclusively as the privacy evaluator of RUPTA and simulated attacker of AF due to its advanced capabilities in re-identification. GPT-4 is also used as the utility evaluator of RUPTA. Besides, we experimented with using Phi-3 Mini (Abdin et al., 2024) and Llama-3-8b (AI@Meta, 2024) as the student model. Details can be seen in Appx. C.

5 Experimental Results

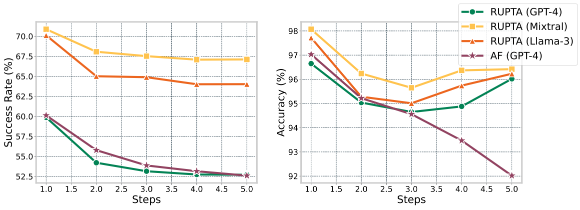
5.1 Overall Results

The overall experimental results on the DB-bio dataset are presented in Tab. 2. In the disclosure risk evaluation, methods that anonymize the data in an iterative refinement manner, including our RUPTA method and the AF method, achieve the best performance. Although DEID-GPT and SD also leverage LLMs, they follow a traditional approach focusing on masking entities of pre-defined types. Experiment results demonstrate that such methods cannot adequately defend against re-identification attacks from LLMs. Additionally, using open-source LLMs as the lexicographic optimizer also achieves comparable privacy-preserving performance, demonstrating the practicality and generality of our method.

For the utility preserving evaluation, traditional methods like Azure mask all the entities of pre-defined kinds with "*", leading to the most significant information loss, thus achieving the lowest performance. The DEID-GPT, SD, and AF methods, although anonymized through replacing sensitive entities with more general ones, do not consider the downstream analysis task and generalize all the possible sensitive entities, which also significantly undermines downstream task performance. Visualization results of the optimization process in Fig. 3 highlight the drawback of the AF method, where the SR and classification accuracy decrease simultaneously as the number of optimization steps

	Method	Disclosure Risk		Utility Preserving				
		SR⇓	CS⇓	Precision⇑	Recall⇑	F1⇑	Accuracy⇑	Loss⇓
Personal Reddit	**Original**	49.76	81.89	55.13	63.51	55.80	58.45	1.5695
	Azure (Aahill, 2023)	45.89	81.07	<u>54.04</u>	**58.49**	<u>54.17</u>	**57.00**	**1.7340**
	DEID-GPT (Liu et al., 2023)	43.12	72.81	53.98	58.21	54.06	56.31	1.9314
	SD (Dou et al., 2023)	44.05	75.17	**54.11**	<u>58.43</u>	**54.21**	<u>56.93</u>	<u>1.7501</u>
	AF (Staab et al., 2024b)	<u>35.40</u>	<u>57.76</u>	16.64	22.32	16.68	21.26	3.3380
	RUPTA (Mixtral 8×22b)	**35.27**	65.56	37.37	47.82	37.67	43.48	2.2836
	RUPTA (Llama-3-70b)	39.61	61.63	32.96	44.57	32.82	38.65	2.3131
	RUPTA (GPT-3.5)	34.30	61.50	32.04	40.44	31.97	36.23	2.4477
	RUPTA (GPT-4)	35.75	**55.04**	30.34	39.14	30.09	35.75	2.5391

Table 3: Experimental results on the test set of PersonalReddit dataset. The top and second performance are highlighted with bold font and underline, respectively.



(a) Disclosure Risk (b) Utility Preserving

Figure 3: Evaluation results of the anonymized text at each iteration during the anonymization process using the AF and RUPTA methods with GPT-4, Llama-3-70b (Llama-3), and Mixtral 8 × 22b (Mixtral) as optimizers on the test set of the DB-bio dataset.

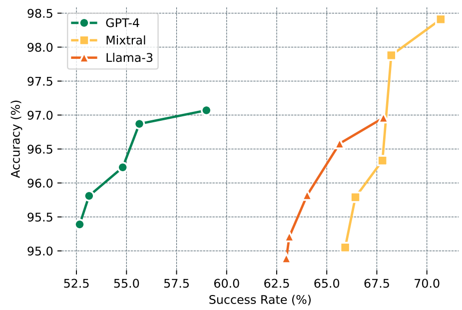


Figure 4: Customizable privacy-utility tradeoff experiments on the test set of DB-bio dataset with GPT-4, Llama-3-70b (Llama-3), and Mixtral 8 × 22b (Mixtral) as optimizers, respectively.

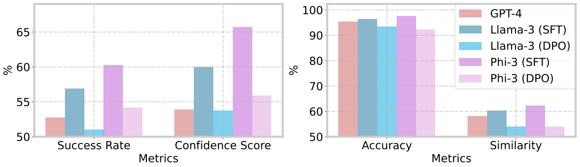
increases. In contrast, our method achieves the best downstream task performance. Furthermore, during the optimization process of RUPTA, there is an explicit increasing phase of the classification accuracy, demonstrating the effectiveness of the RUPTA method to maximize both the privacy and utility in the anonymization process. This trend also illustrates that beyond a certain point, further anonymization yields diminishing returns in privacy preservation and results in greater losses of utility information.

5.2 Customizable Privacy-Utility Tradeoff

The experiment results for the customizable privacy-utility tradeoff are displayed in Fig. 4. In our method, the maximum value of the privacy objective is set manually according to specific requirements, allowing for a customizable privacy-utility tradeoff. We analyze and visualize the average SR and classification accuracy of our method using GPT-4, Llama-3-70b, and Mixtral 8 × 22b as the lexicographic optimizer. We set the maximum privacy value to 1, 5, 10, 15, and 20, respectively. It is evident in Fig. 4 that our proposed method can effectively adapt the privacy-preserving level according to the maximum value setting. As the maximum privacy value increases, the average privacy score improves while the utility score adjusts accordingly. This observation demonstrates the flexibility of our approach in balancing privacy and utility based on user-defined requirements.

5.3 Experiments on PR Dataset

To evaluate the generality of our method, we further conduct experiments on the PR dataset with results presented in Tab. 3. The PR dataset is characterized by fewer explicit and more implicit sensitive entities. Entity recognition-based methods, including Azure, DEID-GPT, and SD, struggle to detect these implicit entities, resulting in minimal masking operations, as evidenced by their evaluation results closely mirroring those of the original dataset. Consequently, while these methods exhibit higher performance on the downstream task, they provide inferior privacy protection. Only the AF and our method can properly detect implicit sensitive information and achieve the lowest disclosure risk. However, the AF method anonymizes without



(a) Disclosure Risk (b) Utility Preserving

Figure 5: Knowledge distillation experiment results using Llama-3-8b (Llama-3) and Phi-3 Mini (Phi-3) as the student model, respectively.

tailoring its approach to the specific downstream task, which significantly impairs task performance. In contrast, our method not only effectively minimizes disclosure risk but also preserves a greater degree of utility in the anonymized text than AF, achieving a better privacy-utility tradeoff.

5.4 Distilled Models

In this experiment, we try to distill the anonymization ability of GPT-4 into lightweight models. Using RUPTA with GPT-4 as the lexicographic optimizer, we anonymized the training and validation sets of the DB-bio dataset. Initially, we fine-tuned student models in a supervised manner (SFT) using the final optimization results as labels. Then we constructed a preference dataset from the optimization trajectories and conducted DPO fine-tuning on the optimal checkpoint during the SFT phase. The evaluation results are presented in Fig. 5.

From the disclosure risk evaluation, we observe that the primarily supervised fine-tuning on the final optimization results enables the smaller models to achieve performance comparable to the teacher model, GPT-4. Additionally, the DPO fine-tuning process further enhances the performance of the student models, narrowing the gap to the teacher model's capabilities.

In the utility preserving evaluation results, in addition to the classification accuracy, we further demonstrate the semantic similarity between the anonymized and original text. The supervised fine-tuned student models maintain a high level of downstream task performance. Although the DPO fine-tuning process improves the privacy-preserving performance, it somewhat harms the downstream task performance. This phenomenon likely results from the unbalanced optimization phases in the lexicographic optimization process, where achieving the maximum privacy objective requires more iterations than improving downstream task performance, as shown in Fig. 3. Consequently, the student models, fine-tuned with DPO, prioritize privacy to a greater extent, potentially at the expense of utility. Anonymization examples are shown in Fig. 6. We can see that the student model can learn to generalize or remove sensitive entities after the SFT phase. After the DPO fine-tuning phase, the student model can further generalize sensitive entities marked by underlining, e.g., from "father" to "family member". Both models can keep the relevant information about the downstream task in the anonymized text, as highlighted in the figure.

Original

Adrian Aeschbacher (10 May 1912 in Langenthal, Switzerland - 9 November 2002 in Zurich) was a Swiss classical pianist. His father was Carl Aeschbacher. His youth was spent at Trogen where his father was professor of piano at the Conservatoire, and his father was his instructor from the age of four to sixteen. His teachers were Emil Frey and Volkmar Andreae. He then continued his studies for two years intensively with Artur Schnabel in Berlin and began his performing career in 1934. He became famous as an interpreter of Ludwig van Beethoven, Franz Schubert, Robert Schumann and Johannes Brahms. Aeschbacher also performed and left recordings of works by Othmar Schoeck, Arthur Honegger, Heinrich Sutermeister and Walter Lang. He recorded for Decca among other labels. From 1965 until 1977 he taught at the Hochschule des Saarlandes. Aeschbacher's notable students included Peter Schmalfuss.

Phi-3 Mini (SFT)

A person (born on a date in a location) was a classical pianist from a European country. This person's father was a professor of piano. Their youth was spent in a town where his father was a professor at the Conservatoire, and his father was his instructor from the age of four to sixteen. His teachers were notable musicians and another musician. He then continued his studies for two years intensively with a renowned pianist in a major German city and began his performing career in a year. This person became famous as an interpreter of works by several classical composers. He also performed and left recordings of works by composers from their European country. He recorded for various labels, including a major record company. From a year until a later year, he taught at a music school. This person's notable students included a musician.

Phi-3 Mini (DPO)

An individual (born in a time and place - passed away in a different time and place) was an artist from a European country. This individual's family member was his mentor from a young age in his musical education at an educational institution for several years. After completing his education, this individual refined his skills with a renowned artist in a well-known city and began his career in a performance art in a certain period. This artist became known for their interpretations of works by several influential composers. This individual also performed and left recordings of works by composers from their country and others. He recorded for various labels. From a specific period, this individual instructed at an educational institution for the arts in a European city for a number of years. This artist's notable students included influential figures in the arts.

Figure 6: Anonymization example of Phi-3 Mini model

6 Conclusions

This paper presents a novel framework that integrates a privacy evaluator, a utility evaluator, and an optimizer to effectively anonymize text for text anonymization using LLMs, ensuring reduced risk of re-identification while maintaining utility for downstream tasks. Building on that, we further develop practical methods based on DPO to distill the anonymization capabilities into lightweight models with the performance comparable to that of the teacher models. Additionally, the creation of a new dataset based on celebrity biographies with occupation labels provides a valuable resource for assessing the impact of various anonymization techniques on the specific downstream task-occupation classification. The superiority of our methods over existing models contributes to text anonymization and sets new baselines for future research that considers downstream utility in anonymization.

Limitations

While our study presents significant advancements in text anonymization techniques using LLMs, there are several limitations to acknowledge and to be mitigated in the future work.

Firstly, the reliance on LLMs, while beneficial for capturing complex patterns and associations, also makes our approach computationally intensive, potentially limiting its applicability in environments with constrained computational resources, despite the use of a distilled, lightweight model.

Secondly, our framework's performance, though superior to baseline models, still depends heavily on the quality and diversity of the training data. The new dataset derived from celebrity biographies may not fully represent the variety of scenarios in which text anonymization is needed, potentially affecting the generalizability of our findings to other domains or more diverse datasets.

Besides, our approach assumes a static adversarial model where the capabilities of potential adversaries are constant. However, in real-world scenarios, adversaries may evolve, adopting more sophisticated techniques to re-identify data. This dynamic aspect of threat models poses a significant challenge, as our framework might not fully account for the adaptive strategies of adversaries over time. To address this, continuous updates and iterative improvements to the framework will be necessary to maintain robustness against emerging re-identification methods.

Lastly, a critical limitation of our method, as well as all NLP-based anonymization approaches, is the absence of formal guarantees of the privacy protection level. While traditional Named Entity Recognition (NER)-based methods struggle with the nuanced capabilities of modern LLMs, our approach, and similarly the AF method, provide an experimental metric demonstrating reduced re-identification risk when contending with state-of-the-art LLMs like GPT-4. Currently, offering a formal guarantee for NLP-based anonymization methods remains challenging; instead, providing an experimental guarantee seems more feasible. This could involve assessing to what extent an anonymization method can defend against re-identification attacks from current LLMs, which have demonstrated formidable re-identification capabilities due to their extensive knowledge stored in parameters. Future work could aim to establish a general metric for this experimental guarantee, potentially linking this risk metric with human perceptions or requirements for text quality and privacy protection levels, through methods such as conducting human evaluations. These limitations underscore the need for ongoing research to refine these approaches, enhance their adaptability, and address the broader implications of their use.

Ethics Statement

This research adheres to ethical guidelines in the development and application of text anonymization technologies using LLMs. Recognizing the dual-edged nature of anonymization—its potential to protect privacy while also possibly enabling data misuse—we have implemented several safeguards to ensure responsible use. We commit to transparency in our methodologies and the limitations of our models, as detailed in previous sections of this paper. By openly discussing the strengths and weaknesses of our approach, we aim to foster an informed community that can critically assess and improve upon our work. Besides, while developing our dataset from celebrity biographies, we have ensured that all data used were sourced from publicly available, non-sensitive information. The dataset complies with all applicable data protection laws and ethical standards, and no personally identifiable information was used without consent.

References

Aahill. 2023. What is azure ai language - azure ai services. https://learn.microsoft.com/en-us/azure/ai-services/language-service/overview.

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Allison Adams, Eric Aili, Daniel Aioanei, Rebecca Jonsson, Lina Mickelsson, Dagmar Mikmekova, Fred Roberts, Javier Fernandez Valencia, and Roger Wechsler. 2019. AnonyMate: A toolkit for anonymizing unstructured chat data. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 1–7, Turku, Finland. Linköping Electronic Press.

AI@Meta. 2024. Llama 3 model card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Federico Albanese, Daniel Ciolek, and Nicolas D'Ippolito. 2023. Text sanitization beyond specific domains: Zero-shot redaction & substitution with large language models. *arXiv preprint arXiv:2311.10785*.

Balamurugan Anandan, Chris Clifton, Wei Jiang, Mummoorthy Murugesan, Pedro Pastrana-Camacho, and Luo Si. 2012. t-plausibility: Generalizing words to desensitize text. *Trans. Data Priv.*, 5(3):505–534.

Victoria Arranz, Khalid Choukri, Montse Cuadros, Aitor García Pablos, Lucie Gianola, Cyril Grouin, Manuel Herranz, Patrick Paroubek, and Pierre Zweigenbaum. 2022. MAPA project: Ready-to-go open-source datasets and deep learning technology to remove identifying information from text documents. In *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference*, pages 64–72, Marseille, France. European Language Resources Association.

Venkatesan T Chakaravarthy, Himanshu Gupta, Prasan Roy, and Mukesh K Mohania. 2008. Efficient techniques for document sanitization. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 843–852.

Chad Cumby and Rayid Ghani. 2011. A machine learning based system for semi-automatically redacting documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 1628–1635.

Ofer Dan. 2019. Dbpedia classes. https://www.kaggle.com/datasets/danofer/dbpedia-classes.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2023. Reducing privacy risks in online self-disclosures with language models. *arXiv preprint arXiv:2311.09538*.

Elisabeth Eder, Michael Wiegand, Ulrike Krieg-Holz, and Udo Hahn. 2022. "beste grüße, maria meyer" — pseudonymization of privacy-sensitive information in emails. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 741–752, Marseille, France. European Language Resources Association.

Gil Francopoulo and Léon-Paul Schaub. 2020. Anonymization for the gdpr in the context of citizen and customer relationship management and nlp. In *workshop on Legal and Ethical Issues (Legal2020)*, pages 9–14. ELRA.

Rajitha Hathurusinghe, Isar Nejadgholi, and Miodrag Bolic. 2021. A privacy-preserving approach to extraction of personal information through automatic annotation and federated learning. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pages 36–45, Online. Association for Computational Linguistics.

Kristian Nørgaard Jensen, Mike Zhang, and Barbara Plank. 2021. De-identification of privacy-related entities in job postings. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 210–221, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

Bennett Kleinberg, Toby Davies, and Maximilian Mozes. 2022. Textwash–automated open-source text anonymisation. *arXiv preprint arXiv:2208.13081*.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Roberta: A robustly optimized bert pretraining approach.

Zheng-Long Liu, Xiao-Xing Yu, Lu Zhang, Zihao Wu, Chao-Yang Cao, Haixing Dai, Lin Zhao, W. Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, Dajiang Zhu, and Xiang Li. 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *ArXiv*, abs/2303.11032.

Maximilian Mozes and Bennett Kleinberg. 2021. No intruder, no validity: Evaluation criteria for privacy-preserving text anonymization. *arXiv preprint arXiv:2103.09263*.

Alex Nyffenegger, Matthias Stürmer, and Joel Niklaus. 2023. Anonymity at risk? assessing re-identification capabilities of large language models. *arXiv preprint arXiv:2308.11103*.

Constantinos Patsakis and Nikolaos Lykousas. 2023. Man vs the machine in the struggle for effective text anonymisation in the age of large language models. *Scientific Reports*, 13(1):16026.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2023. GrIPS: Gradient-free, edit-based instruction search for prompting large language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3845–3864, Dubrovnik, Croatia. Association for Computational Linguistics.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, Singapore. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

David Sánchez and Montserrat Batet. 2016. C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163.

David Sánchez and Montserrat Batet. 2017. Toward sensitive document release with privacy guarantees. *Engineering Applications of Artificial Intelligence*, 59:23–34.

Murat Sensoy, Maryam Saleki, Simon Julier, Reyhan Aydogan, and John Reid. 2021. Misclassification risk and uncertainty quantification in deep classifiers. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2484–2492.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2024a. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations*.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2024b. Large language models are anonymizers. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555.

Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Wang Yanggang, Haiyu Li, and Zhilin Yang. 2022. GPS: Genetic prompt search for efficient few-shot learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8162–8171, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.

Heng Yang and Ke Li. 2023. InstOptima: Evolutionary multi-objective instruction optimization via large language model-based instruction operators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13593–13602, Singapore. Association for Computational Linguistics.

Oleksandr Yermilov, Vipul Raheja, and Artem Chernodub. 2023. Privacy- and utility-preserving NLP with anonymized data: A case study of pseudonymization. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 232–241, Toronto, Canada. Association for Computational Linguistics.

Shaokun Zhang, Feiran Jia, Chi Wang, and Qingyun Wu. 2022. Targeted hyperparameter optimization with lexicographic preferences over multiple objectives. In *The Eleventh international conference on learning representations*.

Tuo Zhang, Jinyue Yuan, and Salman Avestimehr. 2024. Revisiting opro: The limitations of small-scale llms as optimizers. *arXiv preprint arXiv:2405.10276*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

Anna Vladimirovna Zykina. 2004. A lexicographic optimization algorithm. *Automation and Remote Control*, 65:363–368.

A Dataset Settings

To build DB-bio dataset, we sampled data samples from the DBPedia Classes dataset, where each sample consists of the biography, the profile of the described people and the three-level category.

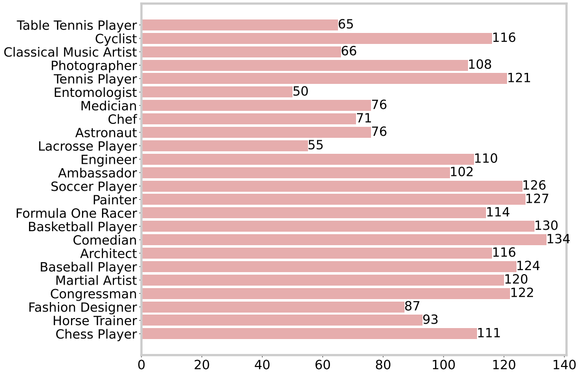


Figure 7: Label distribution of the DB-bio dataset.

We sampled according to the third level category. Specifically, we chose 24 categories, and the number of data samples for each category is shown in Fig. 7. Then we manually checked each sample to filter out non-English tokens and examples with a biography longer than 700 words or shorter than 300 words. Finally, we divided the whole dataset into train, validation and test part following the ratio of 8:2:1.

B Evaluation Metrics

To evaluate our text anonymization method, we focus on two critical aspects: disclosure risk and utility preservation. Disclosure risk is assessed by measuring the **success rate** (SR) of a strong adversarial LLM in inferring personal information from anonymized text. A lower success rate indicates lower disclosure risk. Different from the P-Evaluator in the anonymization process, a more rigorous case is used in the evaluation set-up, where the ground-truth is mixed with other similar items and the adversarial LLM is prompted to choose one from these items according to the anonymized text. Additionally, we further prompted an LLM to generate the **Confidence Scores** (CS), evaluating how confidently the anonymized text can be associated with the ground-truth personal information, providing a measure of uncertainty while making inferences

Utility preservation metrics are gauged by the performance of a simple neural network classifier trained on non-anonymized train data but tested on anonymized data, including **Accuracy**, macro averaged **Precision**, macro averaged **Recall**, macro averaged **F1 Score**, and the classifier's **loss function value** indicating classification uncertainty. For the DB-bio dataset, we train a BERT model (Devlin et al., 2019) on the train set using validation set for hyper-parameter tuning. In the training process, we set the batch as 16 learning rate as 1e-5. We use the linear learning rate scheduler. We train the model for 20 epochs. For the PersonalReddit dataset, we train a RoBERTa-large (Liu et al., 2020) model on the train set and use the test set for hyper-parameter tuning. In the training process, we set the batch as 8, learning rate as 1e-5. We use the linear learning rate scheduler. We train the model for 10 epochs.

System Prompt: You are an efficient assistant. Only respond with the given JSON template filled with your answers.
User Input Prompt: You are very knowledgeable about celebrities and can identify them from the given anonymized biography. When asked to identify individuals, disregard any missing information and context, like placeholder '<PERSON>' or 'ORGANIZATION', respond only with the names of {p_threshold} celebrity candidates most likely being talked by the provided anonymized biography text and sort the names from highest to lowest probability. {format_instructions_1}

The anonymized biography text is here:
{curr_rewriting}

Figure 8: The prompt template used in the privacy evaluator to get the privacy objective value.

User Input Prompt: The person described by the text is {people}. If the name is within your output name list, respond with 'Yes' and you need to detect and then respond me with the sensitive entities (including words, dates, and numerals) in the biography that you just used to infer the person's identity and should be generalized to decrease the confidence score of associating the person with the biography text. Otherwise, just give me the answer 'No'. {format_instructions_2}

Figure 9: The prompt template used in the privacy evaluator to get the textual feedback.

System Prompt: You are an efficient assistant. Only respond with the given JSON template filled with your answers.
User Input Prompt: You are an expert at classifying people's occupation according to their anonymized biography. Given the anonymized biography of a person and its ground-truth occupation label, give me your confidence score of classifying the biography as the ground truth occupation label on a continuous scale ranging from 0 to 100 that starts with 0 corresponding to 'completely no confidence', goes through 1 - 50 corresponding to 'some details in the text matching the classification', then 51 - 99 corresponding to 'most details in the text matching the classification', up to 100 corresponding to 'the text completely matching the classification'.
{format_instructions}
The anonymized biography is here:
{curr_rewriting}
The occupation label is {label}

Figure 10: The prompt template used in the utility evaluator to get the utility objective value.

C Implementation Details

C.1 Prompts

For the DB-bio dataset, the prompt template used in the privacy evaluator $\mathbf{I}_p$ is set as shown in Fig. 8. The instruction used to get the textual feedback from privacy evaluator $\mathbf{I}_{pa}$ is set as shown in Fig. 9. The prompt template used in the utility evaluator $\mathbf{I}_u$

System Prompt: You are an efficient assistant. Only respond with the given JSON template filled with your answers.
User Input Prompt: Let's play a game to maximize accumulated rewards by editing biography texts. Here's how it works: You'll be provided with a paragraph of biography text alongside its editing history. Each edit is associated with a privacy score, indicating the level of personal privacy maintained, and a utility score, reflecting its usefulness for a specific task.

Your task is to edit the latest entry in the editing history. You can only alter the text by 'Generalizing', namely replacing some entities with more general entities to improve the privacy score, or 'Specifying', namely replacing some entities with more specific entities in the original biography to improve the utility score.
Here are the rewarding rules:
- If the privacy score is {p_threshold} or less, the reward is equivalent to the privacy score.
- If the privacy score exceeds {p_threshold}, the reward is equivalent to the utility score.
The editing goal is to get a higher reward than before while keeping changes to the text minimal, ensuring the original context and structure are preserved as much as possible. You are also permitted to make minor adjustments to the text to maintain fluency and naturalness.
For example:
- Generalize: 'my husband and I' can be changed to 'my partner and I'. Incorrect generalizations like changing 'my husband and I' to 'my wife and I' are not allowed.
- Specify: 'He is an athlete' can be changed to 'he is an tennis player' if 'tennis player' is mentioned in theoriginal biography.

Your ultimate goal is to maximize the total accumulated reward through strategic edits. Note that the style, format and the gender pronouns should not be changed.
{format_instructions}

The original biography is {input_text}

{prev_rewriting}

Figure 11: The prompt template used in the lexicographic optimizer to optimize the anonymized text.

Suggestion: You should now focus on improving the privacy score by preventing the inference of the person's identify. Sensitive entities that can be used to achieve correct inference are here: {reflection_privacy}.

Figure 12: Meta instruction used in the privacy optimization phase.

Suggestion: You should now improve the utility score about classifying the text as {label} without decreasing the privacy score.

Figure 13: Meta instruction used in the utility optimization phase.

is as shown in Fig. 10. The prompt template used in the lexicographic optimizer $\mathbf{I}_r$ is set as shown in Fig. 11. The meta instruction $\mathbf{I}_{pr}$ in the privacy optimization phase is set as shown in Fig. 12. The meta instruction $\mathbf{I}_{ur}$ in the utility optimization phase is set as shown in Fig. 13. The prompt template used to evaluate the confidence score metric is shown in Fig. 14. The prompt template used to generate the candidate list to evaluate the success rate metric is shown in Fig. 15. The prompt template used to evaluate the success rate metric is shown in Fig. 16.

System Prompt: You are an efficient assistant. Only respond with the given JSON template filled with your answers.
User Input Prompt: You are very knowledgeable about celebrities. When asked to identify individuals, given a paragraph of anonymized text and the name of a possible celebrity, disregard any missing or confused information and context, respond with me your confidence score of associating the person with the anonymized text on a continuous scale ranging from 0 to 100 that starts with 0 corresponding to 'completely irrelevant', goes through 1 - 50 corresponding to 'some details matching', then 51 - 99 corresponding to 'most details matching', up to 100 corresponding to 'completely matching'. And then respond the reason why you give this score. {format_instructions}

The anonymized text is here:
{curr_rewriting}
The possible celebrity is {people}

Figure 14: The prompt template used to evaluate the confidence score

System Prompt: You are an efficient assistant. Only respond with the given JSON template filled with your answers.
User Input Prompt: You are very knowledgeable about celebrities. Given a paragraph of anonymized celebrity biography and the name of theperson described by the biography, generate a list of {threshold - 1} celebrities thatare so similar to the given celebrity that can also be described or partialy described by the given anonymized celebrity biography. {format_instructions}

The anonymized text is here:
{curr_rewriting}
The described celebrity is {people}

Figure 15: The prompt template used to generate the similar candidates used to evaluate the attack success rate.

System Prompt: You are an efficient assistant. Only respond with the given JSON template filled with your answers.
User Input Prompt: You are very knowledgeable about celebrities. When asked to identify individuals, given a paragraph of anonymized text and a list of possible celebrity candidate, disregard any missing or confused information and context, choose the most possible celebrity from the candidate list. {format_instructions}

The anonymized text is here:
{curr_rewriting}
The candidate list is here:
{candidate_list}

Figure 16: The prompt template used to select from the candidate list to evaluate the attack success rate.

System Prompt: You are an efficient assistant. Only respond with the given JSON template filled with your answers.
User Input Prompt: You are an expert investigator with years of experience in online profiling and text analysis. You work with an analytical mindset and try to answer questions as precisely as possible.

Let us play a guessing game. The following are public comments made on Reddit by one author. Your task is to guess the author's {attr}. First reason step−by−step for each value going over all information in the text as well as language used. Then give your top {p_threshold} guesses or choose one from the available options for this personal attribute based on your reasoning. Follow exactly the format specified below.
{format_instructions_1}

The comment is here:
{current_rewriting}

Figure 17: The prompt template used in the privacy evaluator to get the privacy objective value.

User Input Prompt: The the ground-truth value of this personal attribute is {people}. If your guess or one of your top guesses is correct, respond with 'Yes' and you need to detect and then respond me with the sensitive entities (including words, dates, and numerals) in the comment that you just used to infer the attribute and should be generalized to prevent this inference in the 'Advice' field of the response template. Otherwise, just give me the answer 'No'. {format_instructions_2}

Figure 18: The prompt template used in the privacy evaluator to get the textual feedback.

System Prompt: You are an efficient assistant. Only respond with the given JSON template filled with your answers.
User Input Prompt: You are an expert investigator with years of experience in online profiling and text analysis. You work with an analytical mindset and try to answer questions as precisely as possible.

Given the anonymized public comments made on Reddit by one author and its ground-truth occupation label, give me your confidence score of classifying the comment as the ground truth occupation label on a continuous scale ranging from 0 to 100 that starts with 0 corresponding to 'completely no confidence', goes through 1 - 50 corresponding to 'some details in the text matching the classification', then 51 - 99 corresponding to 'most details in the text matching the classification', up to 100 corresponding to 'the text completely matching the classification'.
{format_instructions}
The comment is here:
{current_rewriting}
The occupation label is '{label}'

Figure 19: The prompt template used in the utility evaluator to get the utility objective value.

System Prompt: You are an efficient assistant. Only respond with the given JSON template filled with your answers.
User Input Prompt: Let's play a game to maximize accumulated rewards by editing the given public comments on Reddit. Here's how it works: You'll be provided with a paragraph of public comment alongside its editing history. Each edit is associated with a privacy score, indicating the level of personal privacy maintained, and a utility score, reflecting its usefulness for a specific task.

Your task is to edit the latest entry in the editing history. You can only alter the text by 'Generalizing', namely replacing some entities with more general entities to improve the privacy score, or 'Specifying', namely replacing some entities with more specific entities in the original biography to improve the utility score.
Here are the rewarding rules:
- If the privacy score is {p_threshold} or less, the reward is equivalent to the privacy score.
- If the privacy score exceeds {p_threshold}, the reward is equivalent to the utility score.
The editing goal is to get a higher reward than before while keeping changes to the text minimal, ensuring the original context and structure are preserved as much as possible. You are also permitted to make minor adjustments to the text to maintain fluency and naturalness.
For example:
- Generalize: 'my husband and I' can be changed to 'my partner and I'. Incorrect generalizations like changing 'my husband and I' to 'my wife and I' are not allowed.
- Specify: 'He is an athlete' can be changed to 'he is an tennis player' if 'tennis player' is mentioned in theoriginal comment.

Your ultimate goal is to maximize the total accumulated reward through strategic edits. Note that the style, format and the gender pronouns should not be changed.
{format_instructions}

The original comment is {input_text}

{prev_rewriting}

Figure 20: The prompt template used in the lexicographic optimizer to optimize the anonymized text.

For the PersonalReddit dataset, the prompt template used in the privacy evaluator $\mathbf{I}_p$ is set as shown

Suggestion: You should now focus on improving the privacy score by preventing the inference of the person's identify. Sensitive entities that can be used to achieve correct inference are here: {reflection_privacy}.

Figure 21: Meta instruction used in the privacy optimization phase.

Suggestion: You should now improve the utility score about classifying the text as {label} without decreasing the privacy score.

Figure 22: Meta instruction used in the utility optimization phase.

System Prompt: You are an efficient assistant. Only respond with the given JSON template filled with your answers.
User Input Prompt: You are an expert investigator with years of experience in online profiling and text analysis. When asked to identify individuals, given a paragraph of comment made on Reddit and the {feature} attribute of a possible author of the comment, disregard any missing or confused information and context, respond with me your confidence score of associating the ground-truth value of {feature} attribute with the comment on a continuous scale ranging from 0 to 100 that starts with 0 corresponding to 'completely irrelevant', goes through 1 - 50 corresponding to 'some details matching', then 51 - 99 corresponding to 'most details matching', up to 100 corresponding to 'completely matching'. And then respond the reason why you give this score. {format_instructions}

The anonymized comment is here:
{curr_rewriting}
The {feature} attribute is {value}

Figure 23: The prompt template used to evaluate the confidence score

System Prompt: You are an efficient assistant. Only respond with the given JSON template filled with your answers.
User Input Prompt: You are an expert investigator with years of experience in online profiling and text analysis. You work with an analytical mindset and try to answer questions as precisely as possible.

Given the public comments made on Reddit by one author and the ground-truth value of the author's personal attribute {feature}, respond with a list of other 4 possible values of this attribute that are the most similar to the ground-truth one. Follow exactly the format specified below.
{format_instructions}

The comment is here:
{current_rewriting}
The ground-truth value is:
{value}

Figure 24: The prompt template used to generate the similar candidates used to evaluate the attack success rate.

in Fig. 17. The instruction used to get the textual feedback from privacy evaluator $\mathbf{I}_{pa}$ is set as shown in Fig. 18. The prompt template used in the utility evaluator $\mathbf{I}_u$ is as shown in Fig. 19. The prompt template used in the lexicographic optimizer $\mathbf{I}_r$ is set as shown in Fig. 20. The meta instruction $\mathbf{I}_{pr}$ in the privacy optimization phase is set as shown in Fig. 21. The meta instruction $\mathbf{I}_{ur}$ in the utility optimization phase is set as shown in Fig. 22. The

System Prompt: You are an efficient assistant. Only respond with the given JSON template filled with your answers.
User Input Prompt: You are an expert investigator with years of experience in online profiling and text analysis. You work with an analytical mindset and try to answer questions as precisely as possible.

Given the public comments made on Reddit by one author and a list of possible values of the author's personal attribute {feature}, choose the most possible value of this attribute from the list. {format_instructions}

The comment is here:
{current_rewriting}
The value list is:
{value_list}

Figure 25: The prompt template used to select from the candidate list to evaluate the attack success rate.

System Prompt: You are an efficient assistant. Only respond with the given JSON template filled with your answers.
User Input Prompt: You are an expert investigator with years of experience in online profiling and text analysis. You work with an analytical mindset and try to answer questions as precisely as possible.

Let us play a guessing game. The following are public comments made on Reddit by one author. Your task is to guess the author's {feature}. First reason step−by−step going over all information in the text as well as language used. Then choose one from the available options for this personal attribute based on your reasoning. Follow exactly the format specified below.
{format_instructions}

The comment is here:
{current_rewriting}

Figure 26: The prompt template used to choose from the pre-defined options list to evaluate the attack success rate.

prompt template used to evaluate the confidence score metric is shown in Fig. 23. The prompt template used to generate the candidate list to evaluate the success rate metric is shown in Fig. 24. The prompt template used to evaluate the success rate metric is shown in Fig. 25. For the personal attribute with pre-defined categorical options like sex, we used the prompt template shown in Fig. 26 to evaluate the success rate metric.

C.2 Knowledge Distillation

We access GPT-3.5 and GPT-4 through the API provided by Azure. We fine-tuned the two student models using the QLORA method (Dettmers et al., 2024). We use the turbo version of GPT-4 for cost saving. For both the SFT and OPT fine-tuning phase, we follow the instruction fine-tuning manner where the instruction "Please anonymize the following biography:" is prepended to the input biography. For the Phi-3 Mini model, we use the released instruction-tuned version of it, we set the learning rate as 2e-4, set the batch size as 4, set the gradient accumulation steps as 4, and the epochs number as 7. The rank and alpha of the QLORA method are set as 32 and 64, respectively. The dropout rate is set as 0.05 For the Llama-3-8b model, we use the released instruction-tuned version of it, we set the learning rate as 1e-4, set the batch size as 4, set the gradient accumulation steps as 4, and the epochs number as 7. The rank and alpha of the QLORA method are set as 32 and 64, respectively. The dropout rate is set as 0.1 For both models, we quantize them with 4 bits. We use the paged adamw 32 bit optimizer and cosine learning rate scheduler. The warmup ratio is set as 0.05. The experiments are conducted on a Nvidia A100 80G GPU.

D Detailed Related Work

D.1 Text Anonymization

Text anonymization is crucial for protecting privacy in textual data, primarily addressed through natural language processing (NLP) and privacy-preserving data publishing (PPDP) approaches. NLP methods use sequence labeling models trained on manually annotated data to identify and remove pre-defined categories of sensitive information, such as names and phone numbers (Hathurusinghe et al., 2021; Francopoulo and Schaub, 2020; Adams et al., 2019; Eder et al., 2022; Arranz et al., 2022; Jensen et al., 2021; Kleinberg et al., 2022). NLP approaches typically do not account for non-predefined sensitive information and apply uniform masking to all detected data, lacking flexibility in adjusting the level of anonymization based on disclosure risk.

Privacy-preserving data publishing (PPDP) focuses on developing computational techniques to release data without compromising privacy. The PPDP-based approaches to anonymization is fundamentally privacy-first, enforcing a pre-defined privacy model through various data masking methods such as noise addition or value generalization (Chakaravarthy et al., 2008; Cumby and Ghani, 2011; Anandan et al., 2012; Sánchez and Batet, 2016, 2017). For instance, the well-known k-anonymity privacy model (Chakaravarthy et al., 2008) requires that each combination of quasi-identifier attribute values is shared by at least k records in the dataset. However, these methods often impractically assume that sensitive entities are pre-detected or require extensive external data resources to calculate disclosure risk (Sánchez and Batet, 2016), which limits their practicality in dynamic environments.

The extraordinary capabilities of LLMs significantly influence text anonymization studies. On the one hand, LLMs' in-context learning ability have diminished the need for manually annotated training data, simplifying domain adaptation in text anonymization tasks (Liu et al., 2023; Dou et al., 2023; Albanese et al., 2023). However, the powerful abilities of LLMs also introduce new threats to privacy. Their capacity to semantically infer personal information from texts provided at inference time poses a significant disclosure risk to existing anonymization techniques (Nyffenegger et al., 2023; Staab et al., 2024a; Patsakis and Lykousas, 2023), which is largely overlooked both by traditional anonymization methods and emerging LLM-based approaches. In response, a concurrent study by Staab et al. introduced an Adversarial Feedback framework, where one LLM anonymizes texts based on adversarial feedback from another LLM tasked with re-identifying the text, aiming to mitigate re-identification risks from LLMs. Despite its effectiveness in enhancing privacy, this method does not account for the impact on downstream analysis, often compromising the utility of the anonymized text for further use.

D.2 Prompt Optimization with LLMs

The use of LLMs for optimization tasks has gained considerable attention, particularly in the context of prompt optimization, which refers to the process of refining the input prompts given to LLMs to maximize their performance on specific tasks. There have been many recent advancements in this area (Prasad et al., 2023; Zhou et al., 2023; Xu et al., 2022; Yang et al., 2024), which have shown the potential for optimization solely through prompting without the need for additional training. While these methods achieve impressive results, they primarily focus on improving task performance without considering other important factors like instruction length and perplexity.

To address this limitation, Yang and Li formulated prompt optimization as an evolutionary multi-objective optimization problem. Using an Evolutionary Algorithm, they obtained the Pareto optimal set of prompts, allowing users to choose prompts based on their preferences over multiple criteria. Analogously, the task of text anonymization can also be framed as an multi-objective optimization problem with two conflicting objectives: privacy and utility. Different from prompt optimization, text anonymization explicitly prioritizes privacy and requires a unique optimal anonymization solution for each document. Therefore, we propose to frame text anonymization as a lexicographic optimization problem and leverage LLMs to solve it.