Robust Utility-Preserving Text Anonymization Based on Large Language Models

Tianyu Yang[1] **Xiaodan Zhu**[1,2] **Iryna Gurevych**[1]
[1]Ubiquitous Knowledge Processing Lab (UKP Lab), Department of Computer Science and Hessian Center for AI (hessian.AI), Technical University of Darmstadt, Germany
[2]Department of Electrical and Computer Engineering & Ingenuity Labs Research Institute, Queen's University, Canada
[1]www.ukp.tu-darmstadt.de [2]xiaodan.zhu@queensu.ca

Abstract

Anonymizing text that contains sensitive information is crucial for a wide range of applications. Existing techniques face the emerging challenges of the re-identification ability of large language models (LLMs), which have shown advanced capability in memorizing detailed information and reasoning over dispersed pieces of patterns to draw conclusions. When defending against LLM-based re-identification, anonymization could jeopardize the utility of the resulting anonymized data in downstream tasks. In general, the interaction between anonymization and data utility requires a deeper understanding within the context of LLMs. In this paper, we propose a framework composed of three key LLM-based components: *a privacy evaluator*, *a utility evaluator*, and *an optimization component*, which work collaboratively to perform anonymization. Extensive experiments demonstrate that the proposed model outperforms existing baselines, showing robustness in reducing the risk of re-identification while preserving greater data utility in downstream tasks. We provide detailed studies on these core modules. To consider large-scale and real-time applications, we investigate the distillation of the anonymization capabilities into lightweight models. All of our code and datasets will be made publicly available at Github[1].

1 Introduction

Privacy protection is a fundamental societal value, enforced in various legal systems such as 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 many others. The recent advancement in AI and large language models (LLMs) presents both challenges and opportunities for privacy protection.

[1]https://github.com/UKPLab/acl2025-rupta

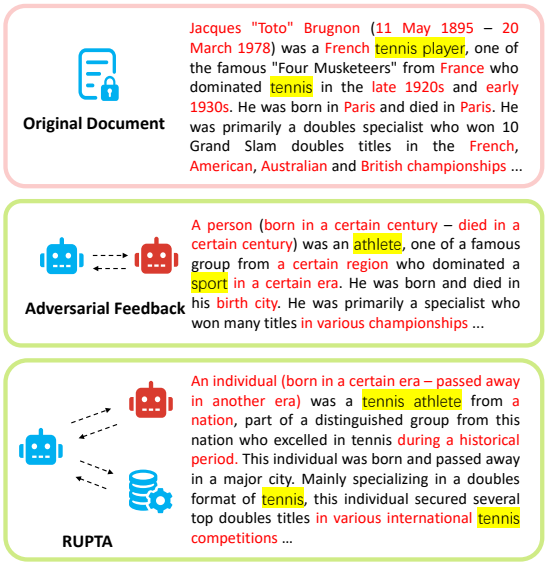


Figure 1: Anonymization examples of the Adversarial Feedback (Staab et al., 2024b) (middle box) and the proposed RUPTA (bottom box) model. The red fonts mark the personally identifiable information. We highlight entities that are critical for our downstream task: occupation classification.

Anonymization is a critical approach to safeguarding private and sensitive information. However, current techniques are vulnerable to disclosure threats from increasingly sophisticated LLMs. For example, recent studies have demonstrated that such models can re-identify private information, even from texts anonymized by advanced methods (Patsakis and Lykousas, 2023; Staab et al., 2024a; Nyffenegger et al., 2024).

The first key challenge and requirement is, therefore, defending against LLM-based re-identification attacks. In combating 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 Figure 1, while the current state-of-the-art (SoTA) method, which conducts anonymization based on iterative refining according to feedback from a simulated

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28922–28941

attacker (Staab et al., 2024b), can defend against re-identification attack well, it may eliminate the information crucial for the downstream task. Existing studies, however, evaluate anonymized text mainly from the perspective of text quality (Dou et al., 2023; Staab et al., 2024b), lacking investigation of the impact on downstream tasks. In general, the interactions between remaining privacy and maintaining utility require a deeper understanding within the context of LLMs, where LLMs' re-identification capacity challenges the safety of existing anonymization models, while if properly utilized, could be leveraged to build more capable anonymization components.

We introduce a 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, 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 evaluations to optimize both objectives until the pre-defined conditions are met. RUPTA outperforms existing baselines based on LLMs, showing robustness in reducing the risk of re-identification while preserving greater data utility in downstream tasks. Note that the privacy protection level can be customized in the proposed framework. Since the anonymization based on LLMs could be time-consuming and resource-intensive, we additionally investigated the distillation of the anonymization capabilities into lightweight models. Our main contributions are summarized as follows:

- To the best of our knowledge, this is the first work to provide comprehensive studies on anonymization and utility in the setup of LLMs, which are crucial for real-world applications.
- We propose a novel framework for text anonymization that is built on the powerful ability of LLMs, consisting of a privacy evaluator, a utility evaluator, and an optimizer component. They work in tandem and show superior performance over the existing models. We provide detailed studies on these core modules. To provide a practical model for real-time environments, we investigate the distillation of anonymization capabilities into smaller 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 downstream tasks. Anonymization results from LLMs are also included to aid future text anonymization research.

2 Related Work

Text Anonymization. To keep privacy when sharing sensitive data, text anonymization serves as a critical alternative to differential privacy-based methods (Feyisetan et al., 2019; Xu et al., 2020; Mattern et al., 2022) and representation learning-based methods (Coavoux et al., 2018) for its high fidelity. This task is primarily addressed through techniques from natural language processing (NLP) and privacy-preserving data publishing (PPDP) (Lison et al., 2021). NLP techniques generally employ sequence labeling models, which are trained on manually annotated datasets to identify and obscure predefined categories of sensitive entities such as names and phone numbers (Hathurusinghe et al., 2021; Francopoulo and Schaub, 2020). In contrast, PPDP methods obscure entities based on a disclosure risk assessed through a privacy model, which is defined by domain experts—examples include C-sanitize (Sánchez and Batet, 2016, 2017). However, most existing studies neglect the utility of anonymized text for downstream tasks or perform post-anonymization evaluations focused on text quality (Yermilov et al., 2023; Staab et al., 2024b), compromising the flexibility of strategies that are able to consider both privacy and utility.

Furthermore, commonly used datasets (Lebret et al., 2016; Pilán et al., 2022) often lack labels for specific downstream tasks, making it difficult to assess the impact of anonymization operations on them.

Anonymization in the Context of LLMs. Significant advancements in LLMs have introduced both challenges and opportunities for text anonymization. Prior to the advent of LLMs, many studies on text anonymization (Sánchez and Batet, 2016, 2017) focused on potential re-identification risks posed by adversaries that could exploit extensive external background knowledge, such as information available on the Web. These studies typically relied on relatively simple re-identification methods that have low attack success rates, such as analyzing the (co-)occurrence counts of terms on

the web or examining lexical and taxonomic relationships. With the rapid development of LLM's reasoning abilities, Staab et al. (2024a) and Patsakis and Lykousas (2023) for the first time demonstrated that LLMs can re-identify anonymized text with remarkable accuracy and speed. Staab et al. (2024b) attempted to harness the capabilities of LLMs to defend against re-identification attacks facilitated by LLMs themselves through deploying an simulated adversarial LLM and using its feedback to inform the anonymization process. Based on this work, Frikha et al. (2024) conducted anonymization by introducing synthetic information to mislead the attacker and introduced an early stopping mechanism to mitigate the deterioration of the utility of the anonymized text. Additionally, Dou et al. (2023) explored interactive anonymization methods that involve fine-tuning LLMs. While these approaches have demonstrated promising performance, their impact on the utility of anonymized text for downstream tasks remains underexplored.

3 Our Methods

We present the Robust Utility-Preserving Text Anonymization (RUPTA) framework, aiming to protect the privacy of sensitive text against the re-identification attack from LLMs while maintaining its utility for downstream tasks.

The overview of RUPTA is depicted in Figure 2. Given a span of text $\boldsymbol{x}_0$, RUPTA iteratively refines the text to optimize the privacy and utility objectives simultaneously. At iteration $t+1$, the previously anonymized text $\boldsymbol{x}_t$ is taken as input, as shown in the bottom left of the figure. The privacy evaluator (P-Evaluator) analyzes $\boldsymbol{x}_t$ to determine its privacy protection level based on the ground-truth personal information y and then provides 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. As shown in the top-right part of the figure, feedbacks from both evaluators are consequently used by the optimizer to refine the text using available editing operations, producing the updated text $\boldsymbol{x}_{t+1}$. Details of the involved instructions can be found in Appendix G.2.

3.1 Problem Formulation

We formulate anonymization as a multi-objective optimization problem, taking into account two objectives: privacy protection and utility of anonymized text. Specifically, this is formulated as Lexicographic Optimization (LO) task (Zykina, 2004) in which we order the two objectives by giving privacy a higher priority—the primary objective is to maximize the level of privacy protection, ensuring that sensitive information is well-protected against re-identification risks. The secondary objective is to preserve as much useful information as possible in the anonymized text for analytical tasks. The optimization problem can be formally expressed as follows:

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

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, RUPTA takes an iterative method based on LLMs to generate, evaluate, and optimize the anonymized text. The details of our LO module are discussed below in Section 3.4.

3.2 The P-Evaluator

The role of our *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. 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 depict the detailed process of P-Evaluation in Algorithm 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 Algorithm 1, where $||$ denotes concatenation. This step generates top-K re-identification 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

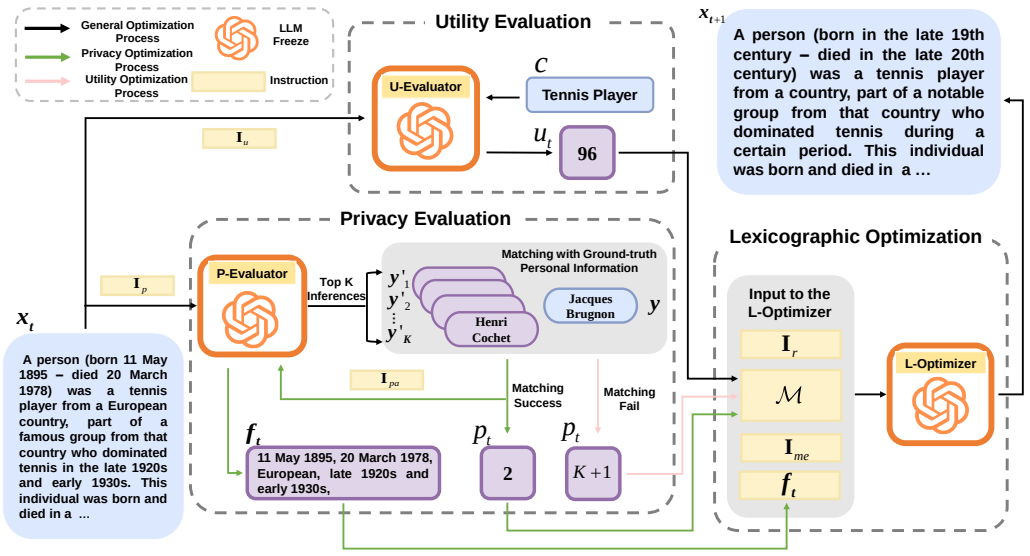


Figure 2: An overview of the proposed RUPTA framework. $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$ denote the input and output text in one iteration; y denotes the ground-truth personal information; and $\boldsymbol{f}_t$, $[y_i']_1^K$ and p_t are the inference feedback, inferred personal information from P-Evaluator and the value of the privacy objective. The ground-truth downstream task label is denoted as c, while u_t is the value of the utility objective. $\mathcal{M}$ denotes the history optimization results. $\mathbf{I}_u, \mathbf{I}_p, \mathbf{I}_r, \mathbf{I}_{me}$ and $\mathbf{I}_{pa}$ are the prompts used for each component of the method.

Algorithm 1 Privacy Objective Evaluation f_p

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

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

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 p_t as $K+1$, representing the maximum score for the privacy objective.

The score p_t quantifies the privacy risk associated with the anonymized text, while the textual feedback $\boldsymbol{f}_t$ offers qualitative insights, guiding the optimizer to better obscure identifiable information. Note that the value of K serves as a parameter that adjusts the sensitivity of the privacy evaluation, with higher values indicating a more inclusive search for potential privacy breaches, thus facilitating a customizable privacy protection level.

3.3 The U-Evaluator

The *Utility Evaluator* (*U-Evaluator*) is designed to ensure that the anonymized text retains its utility for downstream analytical tasks, a crucial consideration for practical applications. It analyzes the anonymized text $\boldsymbol{x}_t$, assessing its effectiveness in supporting the ground-truth label 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.

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

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

which quantifies the evaluator's uncertainty about whether $\boldsymbol{x}_t$ can be correctly related into the ground truth label c, reflecting the degree to which key utility information is preserved. Note that RUPTA is flexible in that the U-Evaluator can be instantiated with the actual model employed in the downstream task. For example, in applications where anonymized text is intended to be used for sentiment analysis (SA), the U-Evaluator can be instantiated with an SA model. The utility score u_t can be calculated using the logit of the ground-truth label following traditional uncertainty quantification methods (Sensoy et al., 2021).

3.4 The Optimizer

Lexicographic optimization (LO) is a special case of multi-objective optimization problems where

multiple conflicting objectives are to be optimized simultaneously. Again, the general objective of LO has been given above in Section 3.1, where the sub-objectives are ranked in order of importance, enabling the prioritization of the more critical objectives. The LO solver is often built on sequential optimization methods (Zykina, 2004; Zhang et al., 2022). In our text anonymization problem, we prioritize privacy over utility.

RUPTA employs an LLM as the lexicographic optimizer by prompting it iteratively to acquire solutions based on the history of optimization results and objective evaluation results. The overall prompt consists of the pre-defined 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 $\mathcal{M} = \{(\boldsymbol{x}_i, p_i, u_i)|i = 1, 2, ..., t\}$ stores history optimization results and their corresponding privacy and utility objective values.

To ensure that the primary objective of achieving maximum privacy is prioritized over 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 the meta-instruction variable as $\mathbf{I}_{pr}$ that instructs the LLM to further anonymize $\boldsymbol{x}_t$ according to textual feedback $\boldsymbol{f}_t$. The process can be formulated as

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

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.

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

The iterative process continues until either the pre-defined maximum values for both objectives are reached or the maximum number of iterations T is met.

3.5 Distilling Anonymization Ability

Iterative anonymization methods based on LLMs could be time-consuming and resource-intensive. We investigate the sequence-level knowledge distillation (SKD) (Kim and Rush, 2016), where a large model (the teacher) transfers its knowledge to a smaller model (the student). Specifically, we propose to 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 assumed less preferred than the final optimization result. These intermediate and final results form the preference dataset. We fine-tune the student model using the DPO method 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.

4 Experimental Set-up

Datasets. Following previous studies (Staab et al., 2024a; Morris et al., 2022), we evaluate our model on the **DB-bio** and **PersonalReddit** datasets. We investigate the impact of anonymization methods on the occupation classification task, a real-world task involving personal information (De-Arteaga et al., 2019). Occupation classification is critical for applications such as job recommendation platforms and automated hiring tools. In such contexts, anonymizing other personal information without affecting the occupation classification performance is important. In addition, a simple classification task allows us to easily demonstrate the performance of the models through quantitative and qualitative analysis. It helps provide intuitive insights into how anonymization models alter critical words necessary for accurate classification.

- **DB-bio**: Previous anonymization studies (Morris et al., 2022) have been conducted on celebrity data available in Wikipedia. Inspired by that, we sampled celebrity biographies from the DBpedia Classes dataset (Dan, 2019) to build a new DB-bio dataset for our study and future research, where we used the category labels of each celebrity in the DBpedia Classes as the occupation classification label.
- **PersonalReddit** (PR) (Staab et al., 2024a): To further assess our method, we evaluate it on the PersonalReddit dataset where we assume personal attributes like gender and location as

	Method	Disclosure Risk		Utility Preservation				
		SR⇓	CS⇓	Precision⇑	Recall⇑	F1⇑	Accuracy⇑	Loss⇓
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
	IncogniText (Frikha et al., 2024)	58.06	56.28	85.68	89.03	87.32	88.28	0.4842
	AF (Staab et al., 2024b)	<u>52.91</u>	**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**	<u>0.2167</u>
	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**	<u>53.11</u>†	<u>95.58</u>†	<u>96.26</u>†	<u>95.91</u>†	<u>96.02</u>†	**0.1618**†

Table 1: The main experiment results on the test set of the DB-bio dataset. The top and second performances are highlighted with bold font and underlined, respectively. Results of RUPTA (GPT-4) denoted by † are significantly better than that of the AF method under the one-tailed paired t-test ($p < 0.05$).

sensitive information and we use occupations as the labels of the task.

Detailed statistics, including category distributions, are provided in Appendix D.

Evaluation Metrics. The evaluation focuses on two critical aspects: disclosure risk and utility preservation. Disclosure risk is assessed by measuring the **Success Rate** (SR) of a state-of-the-art LLM in inferring personal information from anonymized text. Besides, we prompted the LLM to generate the **Confidence Scores** (CS), evaluating the degree 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 BERT-based classifier finetuned on original train data and tested on the test data that is anonymized by RUPTA and other methods, including **Accuracy**, **Precision**, **Recall**, **F1 Score**, and the classifier's **loss function value** indicating classification uncertainty. More details and discussion about metrics can be seen in Appendix F.

Models in Comparison. We compare RUPTA with the following state-of-the-art models.

- We include an industry-standard text anonymizer from Microsoft **Azure** (Aahill, 2023) as one of our anonymization baselines.
- **AF** (Staab et al., 2024b) is a current state-of-the-art LLM-based method for text anonymization based on an adversarial feedback mechanism.
- **DEID-GPT** (Liu et al., 2023) is a recent model that prompts LLMs to mask out pre-defined types of entities.
- **SD** (Dou et al., 2023) is another state-of-the-art approach prompting LLMs to replace entities with more general concepts.
- **IncogniText** (Frikha et al., 2024) is another LLM-based anonymization method that iteratively rewrites the text according to adversarial feedbacks. Different from **AF**, this method adds synthetic information during rewriting to mislead the attacker and proposes an early stopping mechanism to preserve utility.

The LLMs used in the above models are the state-of-the-art GPT-4 models (Achiam et al., 2023). In addition, we explore the effectiveness of different LLMs for the lexicographic optimizer, including open-sourced Llama-3-70b (AI@Meta, 2024) and Mixtral 8×22b (Jiang et al., 2024), and the proprietary GPT-4 and GPT-3.5 (OpenAI, 2022).

Implementation Details. We use GPT-4 as our backbone LLM to ensure the model is comparable to the baselines. The original non-anonymized dataset is evaluated (**Original**) for reference. For implementation details including those for the distillation models, please refer to Appendix G.

5 Experimental Results

5.1 Overall Performance

The overall experimental results on the DB-bio dataset are presented in Table 1. We can see that in the disclosure risk evaluation, methods that anonymize the data in an iterative refinement manner, including RUPTA, IncogniText and AF, 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 demon-

	Method	Disclosure Risk		Utility Preservation				
		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>
	IncogniText (Frikha et al., 2024)	37.55	60.02	10.19	11.06	10.61	13.47	4.4766
	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**	$30.34^{\dagger}$	$39.14^{\dagger}$	$30.09^{\dagger}$	$35.75^{\dagger}$	$2.5391^{\dagger}$

Table 2: Experimental results on the test set of the PersonalReddit dataset. The top and second performances are highlighted with bold font and underlined, respectively. Results of RUPTA (GPT-4) denoted by † are significantly better than that of the AF method under the one-tailed paired t-test ($p < 0.05$).

strate that such methods are not able to adequately defend against re-identification attacks from LLMs, posing critical concerns to these privacy protection methods. Additionally, we can see that using open-source LLMs to build the optimizer can achieve privacy-preserving performance comparable to closed-source models, demonstrating the generality of our method.

RUPTA achieves the lowest utility loss in utility preservation evaluations compared to other baselines. Anonymization generally reduces data specificity to protect privacy, but this comes at the cost of reduced utility. While the original data provides high utility with minimal privacy protection, full masking maximizes privacy but significantly diminishes utility. It is thus essential to investigate a balance—reducing specificity to safeguard private information while preserving sufficient utility in downstream tasks. Existing baselines, however, only assess utility after anonymization and often fail to maintain this balance, either offering insufficient privacy protection (SD and DEID-GPT) or applying excessive anonymization that compromises utility (e.g., AF and IncogniText). Due to the introduction of distracting synthetic information, IncognitText gets the worst utility-preserving performance. RUPTA respects and achieves super disclosure-risk performances and maintains competitive utility preservation.

We hope this work helps set a new benchmark for text anonymization research in the context of LLMs and under the practical setup of examining both the protection and utility, since without an explicit consideration of the two perspectives, we lose a comprehensive view of the problem.

In Figure 3, the visualization of the evaluation

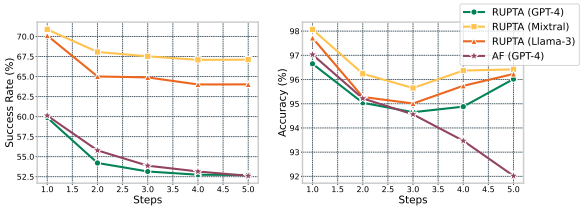
t	1	2	3	4
u_t	43.14	42.31	43.30	44.08

Table 3: Average u_t score during the anonymization process on the PersonalReddit dataset.

results of the optimization process shows that the SR and classification accuracy decrease simultaneously as the number of optimization steps increases. In contrast, our method achieves the best performance in the downstream task. During the optimization process of RUPTA, there is an explicit increase phase of the classification accuracy, demonstrating the effectiveness of the RUPTA method to maximize both privacy and utility in the anonymization process. Similar trends can also be seen from the average u_t score during the anonymization process on the PersonalReddit dataset, as shown in Table 3. This trend illustrates that, beyond a certain point, further anonymization yields diminishing returns in privacy preservation and results in greater losses of utility information. Baseline methods that either ignore the downstream task utility during the anonymization process or only attempt to mitigate its loss afterward often miss this certain point. As a result, they may cause greater utility loss or improve utility at the expense of privacy protection performance. To further demonstrate the effectiveness of RUPTA, we conducted experiments that adapted DEID-GPT and SR as lexicographic optimizers, as detailed in Appendix B.

5.2 Customizable Privacy-Utility Tradeoff

The experiment results for the customizable privacy-utility tradeoff are displayed in Figure 4. In our method, the maximum value of the privacy ob-



(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.

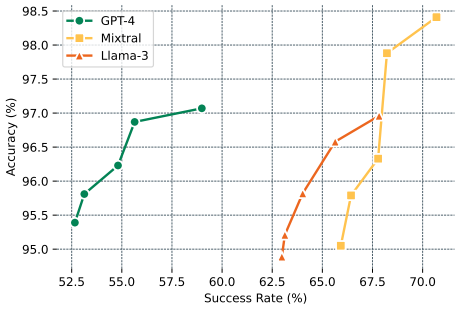


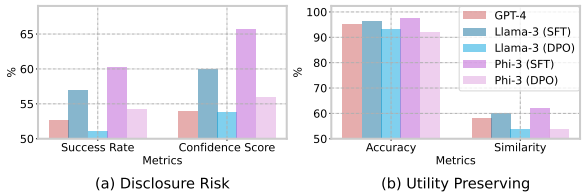
Figure 4: Customizable privacy-utility tradeoff experiments on the test set of the DB-bio dataset with GPT-4, Llama-3-70b (Llama-3), and Mixtral 8×22b (Mixtral) as optimizers, respectively.

jective $K + 1$ 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 Figure 4 that our proposed method can effectively adapt the privacy protection level according to the maximum value setting. As the maximum privacy value increases, the average privacy score improves while the utility score adjusts accordingly. According to Figure 3, the privacy protection level also varies throughout the optimization process. Thus RUPTA can adjust the privacy protection level in a wider range by adjusting both the maximum number of iterations T and K. More experiment results are included in Appendix C.

5.3 Experiments on the PR Dataset

To demonstrate the generality of our method, we further conduct experiments on the PR dataset with results presented in Table 2. The PR dataset is characterized by fewer explicit and more implicit sensitive entities. Entity recognition-based methods,



(a) Disclosure Risk (b) Utility Preserving

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

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 RUPTA can properly detect implicit sensitive information and achieve the lowest disclosure risk. However, the AF method anonymizes without tailoring its approach to the specific downstream task, which significantly impairs task performance. In contrast, RUPTA not only effectively minimizes disclosure risk but also preserves a greater degree of utility in anonymized text than AF.

5.4 Distillation Results

As shown in the computational analysis in Appendix A, anonymization methods based on iterative prompting of LLMs are computationally expensive. We follow the knowledge distillation scheme proposed in §3.5 to distill the anonymization ability of GPT-4 into lightweight models. The evaluation results are presented in Figure 5. More detailed results can be seen in Appendix E.

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 preservation 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. We observed that although the DPO fine-tuning process improves the privacy-preserving performance, it could harm the downstream task performance. We due this to the

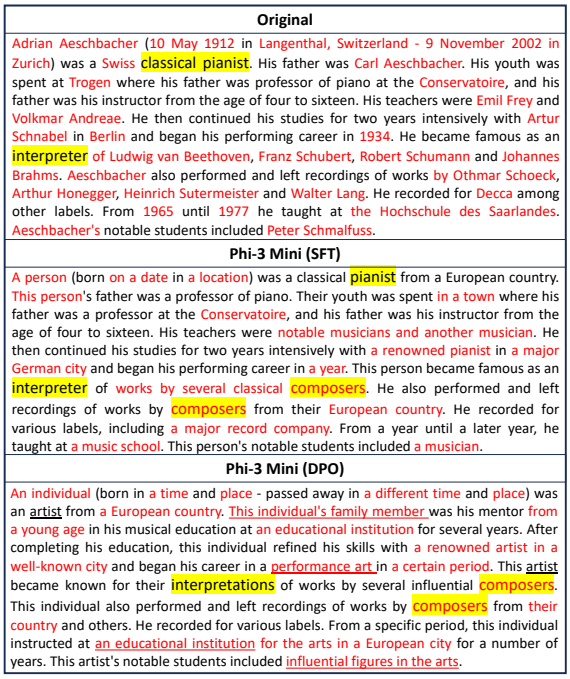


Figure 6: Anonymization example of Phi-3 Mini model

objective of the optimizer in prioritizing privacy over utility, where more optimization steps are utilized to improve the level of privacy protection as shown in Figure 3. The student models have been shown to learn from the optimization history data and prioritize privacy to a greater extent potentially at the expense of utility. Anonymization examples are shown in Figure 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 in the anonymized text for the downstream task, as highlighted in the figure.

5.5 Human Evaluation

To further examine the semantic loss and the quality of generated text, we conducted a human evaluation on a subset of 100 randomly selected examples from the test split of DB-bio, where 3 non-author human subjects rated the anonymization output of Azure, SD, AF and RUPTA using a Likert scale of 1 - 5, on "*how much the original meaning of the non-anonymized text is preserved after anonymization?*" (5 means all information is preserved and 1 means none). In addition, we also asked the human subjects to evaluate the fluency of the anonymized text with the Likert scale of 1 - 5 (5 means the most fluent and 1 means the least). We take an average of the scores of the three human subjects. The results can be seen in Table 4.

Method	Semantic Preservation	Fluency
Azure	3.23	2.09
SD	3.64	3.62
AF	3.78	3.71
RUPTA	3.96	3.68

Table 4: Human evaluation results on the test set of DB-bio.

We can see that Azure introduces the most significant distortion to the original text, as it masks sensitive entities with extensive nonsense special tokens. SD replaces sensitive entities with more general English terms, resulting in better overall semantic preservation and fluency performance. Unlike Azure and SD, which are based on entity-level replacement and adaptation, RUPTA and AF are rewriting-based methods that rewrite the entire passage, which makes them achieve similar or better fluency performance. Compared to the other three methods, RUPTA achieves the best semantic preservation performance, which we believe is due to its consideration of the general semantics of the entire passage during anonymization.

6 Conclusions

This paper presents a novel framework, RUPTA, that integrates a privacy evaluator, a utility evaluator, and an optimizer to effectively anonymize text, 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 a performance comparable to that of the teacher models. Additionally, we create a dataset from celebrity biographies with occupation labels to evaluate the effects of anonymization techniques on specific tasks. The superior performance of RUPTA over existing models and the evaluation setup help establish new baselines for future research that considers downstream task 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 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

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, this work is evaluated on publicly available datasets. While developing our dataset from celebrity biographies, we have ensured that all data used were sourced from publicly available, non-sensitive information.

Acknowledgement

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. We gratefully acknowledge the support of Microsoft with a grant for access to OpenAI GPT models via the Azure cloud (Accelerate Foundation Model Academic Research).

References

Aahill. 2023. What is azure ai language - azure ai services. https://learn.microsoft.com/en-us/azure/ai-services/language-service/overview. Accessed on Jan 12, 2024.

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*, abs/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. Accessed on Apr 20, 2024.

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

Maximin Coavoux, Shashi Narayan, and Shay B. Cohen. 2018. Privacy-preserving neural representations of text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1–10, Brussels, Belgium. Association for Computational Linguistics.

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`. Accessed on Feb 27, 2024.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

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*, abs/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.

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. 2019. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 210–219. IEEE.

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.

Ahmed Frikha, Nassim Walha, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2024. Incognitext: Privacy-enhancing conditional text anonymization via LLM-based private attribute randomization. In *Neurips Safe Generative AI Workshop 2024*.

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*, abs/2401.04088.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Bennett Kleinberg, Toby Davies, and Maximilian Mozes. 2022. Textwash–automated open-source text anonymisation. *ArXiv preprint*, abs/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.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online. 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. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

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 preprint*, abs/2303.11032.

Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan. 2022. Differentially private language models for secure data sharing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4873, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

John Morris, Justin Chiu, Ramin Zabih, and Alexander Rush. 2022. Unsupervised text deidentification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4777–4788, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Alex Nyffenegger, Matthias Stürmer, and Joel Niklaus. 2024. Anonymity at risk? assessing re-identification capabilities of large language models in court decisions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2433–2462, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`. Accessed on Apr 28, 2024.

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.

Nan Xu, Oluwaseyi Feyisetan, Abhinav Aggarwal, Zekun Xu, and Nathanael Teissier. 2020. Differentially private adversarial robustness through randomized perturbations. *arXiv preprint arXiv:2009.12718*.

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*.

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 Computational Cost Analysis

To analyze the computational cost of anonymization method based on iterative prompting LLMs, including AF and RUPTA, we record the average time for anonymizing one paragraph in the DB-bio dataset (AT), average number of prompt tokens (PT) and average number of completion tokens (CT) in Table 5. Each paragraph contains 234 tokens on average in this dataset.

As shown by the results, the computational cost in time and money are both relatively high, which demonstrates the necessity of distilling the knowledge of a large model into a lightweight model to improve the practicality of these methods.

Method	**AT** (s)	**#PT**	**#CT**
RUPTA	76.38	3846.28	697.21
AF	72.57	2979.34	612.01

Table 5: Computational cost analysis of the LLM-based methods.

B Method Transferring

To further demonstrate the effectiveness and practicality of RUPTA, we conducted experiments about adapting previous LLM-based anonymization methods that anonymize by prompting the LLM in a single round including DEID-GPT and SD as the optimizer in RUPTA. Specifically, we append an additional prompt to make them conduct the anonymization according to feedback from the privacy and utility evaluators. We can see from the results in Table 6 that the paradigm of RUPTA can significantly improve the performance of these two baselines. However, due to their being limited to masking or generalizing only entities, their utility preservation performance is still lagged behind RUPTA.

C Customizable Privacy-Utility Tradeoff

In RUPTA, we can manually adjust the privacy-utility tradeoff by setting the maximum of the privacy objective as demonstrated in §5.2. Besides, as shown in Figure 3, the privacy-utility tradeoff is also changed as the number of optimization steps increases. In this section, we explore the effect of the maximum privacy objective in different optimization steps. We implemented RUPTA using Llama-3-70b here. As shown by the results demonstrated in Table 8, the privacy protection level can be actually adjusted in a wider range. Practically, the suitable value of K and T can be empirically chosen by running RUPTA on a validation set and then deploying it in the actual use case. We further conduct experiments to verify the performance of RUPTA is robustness. We list the **SR**, **CS** and **Accuracy** performance of RUPTA on both the validation and test set of DB-bio in Table 7. The result is the average of 5 runs on the validation and test set, respectively. We can see that the performance of RUPTA is consistent across the two subsets of DB-bio.

	Method	Disclosure Risk		Utility Preservation				
		SR⇓	CS⇓	Precision⇑	Recall⇑	F1⇑	Accuracy⇑	Loss⇓
DB-bio	**Original**	100.00	98.45	99.58	99.68	99.61	99.58	0.0422
	DEID-GPT	77.10	79.47	90.82	94.37	92.56	91.22	0.3103
	DEID-GPT*	53.12	53.98	93.01	94.41	93.76	93.83	0.2784
	SD	73.21	73.63	92.27	93.11	92.69	92.96	0.2719
	SD*	**52.43**	54.16	93.98	94.67	94.07	94.10	0.2132
	RUPTA (GPT-4)	52.67	**53.11**	**95.58**	**96.26**	**95.91**	**96.02**	**0.1618**

Table 6: Method transferring experiment results on the test set of DB-bio dataset. The top and second performances are highlighted with bold font and underlined, respectively. denotes the adapted baseline.

Method	SR (s)	CS	Accuracy
RUPTA-test	64.27	64.11	95.92
RUPTA-val	65.16	64.07	95.46

Table 7: Experiment results of RUPTA on the validation and test set of DB-bio.

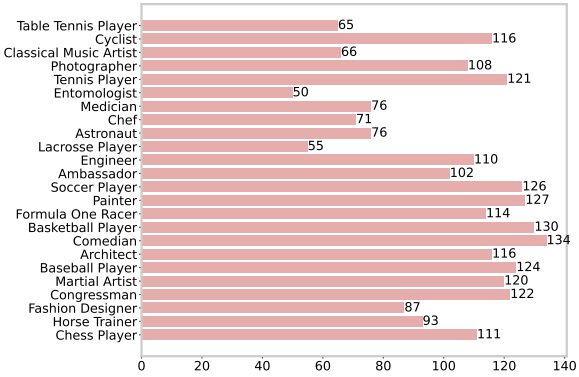


Figure 7: Label distribution of the DB-bio dataset.

Additionally, by comparing the results of **SD** and RUPTA (K=1) with an optimization step of 1, we observe that RUPTA achieves better utility preservation while maintaining nearly the same level of privacy protection.

D Dataset Settings

In this paper, we assume a threat model where the adversaries utilize an LLM pre-trained or fine-tuned on a corpus containing sensitive information to re-identify personal information from the anonymized free-form text (Staab et al., 2024a; Patsakis and Lykousas, 2023). To evaluate our method in this threat model efficiently without collecting a sensitive information dataset and training an adversary LLM upon it, we use the following two datasets:

- **DB-bio**: Many LLMs are pre-trained on the Wikipedia corpus to get the primary knowledge. Thus, we can assume the celebrity information in Wikipedia as personal information to be anonymized and use existing LLMs that have memorized this information as the adversary LLM to attack anonymization methods. 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 lack 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.

- **PersonalReddit** (PR): Due to the rich existence of the celebrity information in the whole pre-train dataset, off-the-shelf LLMs are proficient at guessing the celebrity information from anonymized text. The evaluation performed on the above dataset can only provide an upper bound of the attack success rate. To further validate the generality of our method, we evaluate it on the PR dataset (Staab et al., 2024a) consisting of 525 human-verified synthetic public Reddit comments and the 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 like gender and location that are understood by existing LLMs.

	Method	Disclosure Risk		Utility Preservation				
		SR⇓	CS⇓	Precision⇑	Recall⇑	F1⇑	Accuracy⇑	Loss⇓
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)	<u>52.91</u>	**50.84**	91.20	94.26	91.75	92.02	0.4048
	Optimization step = 1							
	RUPTA (K=1)	72.74	73.69	97.12	98.39	97.67	97.11	0.0867
	RUPTA (K=5)	68.76	70.23	96.34	97.01	96.15	96.82	0.1121
	RUPTA (K=10)	67.32	69.11	96.08	96.56	95.92	96.23	0.1389
	Optimization step = 2							
	RUPTA (K=1)	68.23	69.78	95.12	96.16	95.44	96.09	0.1201
	RUPTA (K=5)	65.02	67.93	94.23	95.07	94.89	94.97	0.1608
	RUPTA (K=10)	64.67	63.11	94.12	95.23	95.03	94.39	0.1820
	Optimization step = 3							
	RUPTA (K=1)	66.18	68.34	95.39	96.11	95.78	96.06	0.1526
	RUPTA (K=5)	65.41	67.14	94.28	95.19	94.88	94.96	0.1599
	RUPTA (K=10)	64.23	67.02	94.09	95.10	94.67	94.29	0.1684

Table 8: Privacy-utility tradeoff experiment results in different optimization steps on the test set of the DB-bio dataset. The top and second performances are highlighted with bold font and underlined, respectively.

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

Table 9: Statistics of experiment datasets.

General statistics of these two datasets can be seen in Table 9.

To build The 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 third-level category. 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 Figure 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 200 words. Finally, we divided the whole dataset into train, validation, and test parts following the ratio of 8:2:1.

E Knowledge Distillation Results

Detailed knowledge distillation experiment results can be seen in Table 10. Row "Phi-3" and "Llama-3" represent the performance of these models without fine-tuning, and the other rows were results copied from Figure 5 of our paper (We extracted the values used to draw Figure 5 and present them numerically in the table below).

Model	SR (s)	CS	Accuracy	Similarity
Phi-3	71.42	74.49	93.22	56.70
-SFT	60.25	65.73	97.48	62.12
-DPO	54.15	55.90	92.15	53.90
Llama-3	69.38	71.56	95.40	58.92
-SFT	56.90	59.97	96.33	60.29
-DPO	51.03	53.78	93.28	54.02

Table 10: Knowledge distillation experiment results.

F 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 asso-

ciated 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 the 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., 2019) model on the train set and use the test set for hyper-parameter tuning. In the training process, we set the batch as 8 and, the learning rate as 1e-5. We use the linear learning rate scheduler. We train the model for 10 epochs.

Due to the difference between the original test set and the anonymized test set, there exists an out-of-distribution (OOD) problem that will affect the performance of our evaluation classifier trained on the original training set. The general viewpoint on OOD issues is that when the mismatch between the training and test datasets is less significant, the neural networks should perform better than otherwise. This is consistent with what we demonstrated in the utility preservation experiments. If the anonymization method keeps the original text as much as possible, then the OOD issue is expected to be less significant , and the performance drop of the neural network classifier on the anonymized test dataset should be less too. As shown in Table 1, named entity recognition (NER)-based methods replaced too many entities and thus got the highest performance drop. While RUPTA achieved the lowest performance drop due to its preservation of the original text. Similarly, as shown in Table 2, NER-based methods suffered from recognizing implicit personally identifiable information, they almost didn't anonymize anything, thus they got the lowest performance drop. Since the goal of non-training metrics is also to compare the original and the anonymized text to evaluate whether the anonymization methods keep the original text as much as possible, we expect the trends on the non-training metrics will be similar to what have observed in the paper. Therefore, we didn't use other non-training metrics to compare the original and anonymized text directly.

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.

G Implementation Details

G.1 Main experiments

For AF and our method, we set the maximum iteration number, namely the value of T, as 5 to make a fair comparison. We set the value of K as 10 to obtain the main results of the experiment. In the experiment, we use the Microsoft Azure platform to access GPT-3.5 and GPT-4, the API version is 2023-05-15. The version of GPT-3.5 is 0301. The version of GPT-4 is 0613.

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.

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.

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.

G.2 Prompts

For the DB-bio dataset, the prompt template used in the privacy evaluator $\mathbf{I}_p$ is set as shown in Figure 8. The instruction used to get the textual feedback from privacy evaluator $\mathbf{I}_{pa}$ is set as shown in Figure 9. The prompt template used in the utility evaluator $\mathbf{I}_u$ is as shown in Figure 10. The prompt template used in the lexicographic optimizer $\mathbf{I}_r$ is set as shown in Figure 11. The meta instruction $\mathbf{I}_{pr}$ in the privacy optimization phase is set as shown in Figure 12. The meta instruction $\mathbf{I}_{ur}$ in the utility optimization phase is set as shown in Figure 13. The prompt template used to evaluate the confidence score metric is shown in Figure 14. The prompt template used to generate the candidate list to evaluate the success rate metric is shown in Figure 15. The prompt template used to evaluate the success rate metric is shown in Figure 16.

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

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.

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.

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.

shown in Figure 22. The prompt template used to evaluate the confidence score metric is shown in Figure 23. The prompt template used to generate the candidate list to evaluate the success rate metric is shown in Figure 24. The prompt template used to evaluate the success rate metric is shown in Figure 25. For the personal attribute with pre-defined categorical options like sex, we used the prompt template shown in Figure 26 to evaluate the success rate metric.

G.3 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 savings. For both the SFT and OPT fine-tuning phases, 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.

H Detailed Related Work

H.1 Text Anonymization

Text anonymization is crucial for protecting privacy in textual data, primarily addressed through natural language processing (NLP) and privacy-preservation 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 are 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 has 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., 2024; 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.

H.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 a 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.