

# A False Sense of Privacy: Evaluating Textual Data Sanitization Beyond Surface-level Privacy Leakage

Rui Xin<sup>1\*</sup> Niloofar Miresghallah<sup>1\*</sup> Shuyue Stella Li<sup>1</sup> Michael Duan<sup>1</sup>  
 Hyunwoo Kim<sup>2</sup> Yejin Choi<sup>3</sup> Yulia Tsvetkov<sup>1</sup> Sewoong Oh<sup>1</sup> Pang Wei Koh<sup>1,4</sup>  
<sup>1</sup>University of Washington <sup>2</sup>NVIDIA <sup>3</sup>Stanford University  
<sup>4</sup>Allen Institute for Artificial Intelligence  
 rx31@cs.washington.edu niloofar@cs.washington.edu

## Abstract

Sanitizing sensitive text data typically involves removing personally identifiable information (PII) or generating synthetic data under the assumption that these methods adequately protect privacy; however, their effectiveness is often only assessed by measuring the leakage of explicit identifiers but ignoring nuanced textual markers that can lead to re-identification. We challenge the above illusion of privacy by proposing a new framework that evaluates re-identification attacks to quantify individual privacy risks upon data release. Our approach shows that seemingly innocuous auxiliary information—such as routine social activities—can be used to infer sensitive attributes like age or substance use history from sanitized data. For instance, we demonstrate that Azure’s commercial PII removal tool fails to protect 74% of information in the MedQA dataset. Although differential privacy mitigates these risks to some extent, it significantly reduces the utility of the sanitized text for downstream tasks. Our findings indicate that current sanitization techniques offer a *false sense of privacy*, highlighting the need for more robust methods that protect against semantic-level information leakage.

## 1 Introduction

It is critical to protect user and patient privacy when sharing data for research and commercial collaborations (Federal Data Strategy, 2020; McMahan et al., 2017). When not properly handled, sensitive data—personal identifiers, location traces, behavioral patterns, etc.—can be exposed through re-identification attacks. During such attacks, adversaries use auxiliary information to analyze released data and infer information about individuals despite sanitization efforts by the data publisher. Re-identification attacks have proven effective for structured datasets (Narayanan & Shmatikov, 2006), where identifiers are clearly defined in tabular format. This effectiveness establishes them as a key method for evaluating data privacy, an approach similar to those found in statistical disclosure control (SDC) guidelines used by the US Census Bureau (Abowd et al., 2023).

Large language models (LLMs) brings new attention to privacy concerns in *unstructured* textual data (Yan et al., 2024). Beyond explicit identifiers, such as personally identifiable information (PII) like names and addresses, text often contains contextual details that can indirectly reveal identity in ways not typically present in structured datasets. *We hypothesize that re-identification risk remains high with textual data even after PII removal.* For example, consider Alice’s record in the sanitized medical dataset shown in Figure 1, where all patient information has been de-identified and PIIs removed. An adversary might possess auxiliary information about Alice, such as her regular social activities or recent job loss, obtainable through personal knowledge or public sources like social media. By exploiting similarities between this external information and entries in the sanitized dataset (Ganta et al., 2008),

---

\*Equal Contribution

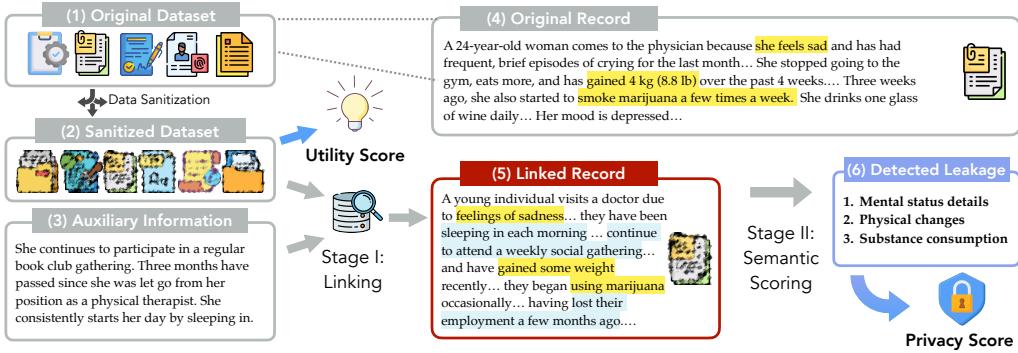


Figure 1: Our privacy evaluation framework overview. For a given sanitization method, we obtain the (2) sanitized dataset by applying the method to the (1) original dataset. In the **linking** stage, we use (3) auxiliary information to find potential matches in the sanitized dataset using a semantic retriever to obtain the (5) linked record. Next, in the **semantic scoring** stage, we analyze the linked record against the corresponding (4) original record to identify *semantic* information leakage. In the linked record, text highlighted in **yellow** indicates detected leakage, and in **cyan** indicates content used for the linking process. The framework calculates *utility scores* to measure the practical value of the sanitized dataset and a *privacy score* based on the detected information leakage.

the adversary could re-identify Alice’s record and then infer sensitive attributes, such as her mental health status or substance use.

Developing appropriate metrics to quantify such re-identification vulnerability in unstructured textual data presents unique challenges. Implementing early theoretical frameworks like adapted k-anonymity (Venkatesan et al., 2008; Lison et al., 2021) and mutual-information approaches (Anandan & Clifton, 2011) remains difficult due to challenges in establishing consistent quasi-identifiers and measuring mutual information. Additionally, standard re-identification attacks assume an one-to-one mapping between original and sanitized documents, which is not available for dataset level sanitization methods such as data synthesis. Moreover, many privacy evaluations rely on lexical metrics such as exact matching (Pilán et al., 2022; Boutet et al., 2025; Carlini et al., 2019; 2022) or ROUGE (Huang et al., 2023), overlooking the fact that the same information can be expressed in different words.

To overcome these challenges, we introduce a new framework for unstructured text that quantifies how much information about an individual can be inferred from sanitized data when combined with auxiliary information (Ganta et al., 2008). The framework employs a two-stage process (Figure 1). The first stage, **linking**, uses a semantic retriever to link auxiliary information with sanitized records. The second stage, **semantic scoring**, measures information leakage by comparing the linked sanitized record to the original private data at the atomic claim level. Simultaneously, we evaluate data utility to assess the tradeoff between privacy and utility.

We evaluate various state-of-the-art sanitization methods on two real-world datasets: MedQA (Jin et al., 2021), with medical notes, and WildChat (Zhao et al., 2024), with AI-human dialogues (Mireshghallah et al., 2024). We compare: (1) *identifier removal techniques*, including commercial PII removal, LLM-based anonymizers (Staab et al., 2024), and sensitive span detection (Dou et al., 2024), and (2) *data synthesis methods* that use GPT-2 fine-tuned on private data, with and without differential privacy (Yue et al., 2023).

*We find that current dataset sanitization methods for text data often provide a false sense of privacy.* Specifically: (1) State-of-the-art PII removal methods operate at a surface level and still exhibit significant leakage, with 74% of original information still inferable. (2) Without differential privacy, synthesized data still exhibits leakage, with 48% of the information re-identifiable. (3) Differentially private (DP) synthesis methods provide the strongest privacy protections but can significantly reduce utility, particularly for complex tasks. On MedQA, DP-synthesized medical notes have lower utility than the degenerate baseline that

removed the original notes. Additionally, text quality dropped by 36%, as evaluated by GPT-4o.

Our evaluation framework establishes that current approaches leave sensitive data vulnerable to re-identification attacks. These results highlight the necessity of developing methods that protect privacy beyond surface-level measures and obvious identifiers and of moving toward semantic evaluation of privacy rather than relying solely on lexical assessment.

## 2 Related Work

**Privacy evaluations of dataset disclosure.** Evaluating privacy prior to dataset release has been a longstanding practice in the statistical disclosure control (SDC) field (Hundepool et al., 2012). This practice spans various domains, including legal, technical, and medical (Bellovin et al., 2019; Garfinkel, 2015; Giuffrè & Shung, 2023). Traditional evaluations have focused on re-identification risks, particularly for tabular data in census or medical settings (Abowd et al., 2023; El Emam et al., 2011). Although there have been attempts to create text anonymization benchmarks (Pilán et al., 2022), the benchmarks primarily address span detection and anonymization and focus on scrubbing methods. Our work examines re-identification and also incorporates modern sanitization methods that generate new data. Additionally, researchers have proposed membership inference attacks to determine whether specific data was used in training and canary attacks to detect the memorization of specific sequences in LMs (Carlini et al., 2019; 2021; 2022). These privacy evaluation approaches differ from our work because they examine models trained on the data rather than on the dataset itself. They depend on access to or assumptions about the model that processed the original text, which is not applicable to all data sanitization techniques. Recent work in the security literature has begun to challenge the perceived safety of synthetic data (Stadler et al., 2022; Yale et al., 2019; Annamalai et al., 2024), raising concerns about its privacy guarantees. However, these investigations primarily focus on simple, low-dimensional tabular or image data and have not extended to unstructured text, leaving a critical gap.

**Data sanitization through removal of identifiers.** To effectively sanitize unstructured text and prevent re-identification, researchers have proposed theoretical privacy frameworks such as adapted k-anonymity (Venkatesan et al., 2008; Lison et al., 2021) and mutual-information-based methods (Anandan & Clifton, 2011). However, implementing these frameworks is often costly and impractical: text can be expressed in many semantically equivalent forms, making it difficult to define consistent quasi-identifiers for k-anonymity or to reliably measure mutual information across varied expressions of the same content. These challenges have led to practical data sanitization approaches that focus on detecting and removing Personally Identifiable Information (PII) (Mendels et al., 2018; Montani et al., 2022) using named entity recognition (NER) followed by masking. Recent work has also explored using LLMs for this task. Staab et al. (2024) developed an iterative prompting method using GPT-4 to achieve implicit attribute removal, moving beyond simple token replacement. Dou et al. (2024) proposed a two-step approach, combining a self-disclosure detection model with an abstraction technique to reduce privacy risks in text data. Other sanitization methods identify sensitive words by prompting an LLM (Zhou et al., 2024). Morris et al. (2022) introduced an unsupervised de-identification method that removes words that could lead to re-identification. However, their method requires a dataset of aligned text and profiles for training, posing unrealistic constraints. These approaches mainly sanitize the dataset by abstracting or removing detected keywords to minimize re-identification and are susceptible to our proposed semantic re-identification attack.

**Data sanitization through synthesis.** To provide untargeted protection at the dataset level, researchers have employed data synthesis (Garfinkel, 2015), occasionally with the assumption that synthesis alone provides some degree of privacy (Liu et al., 2024). For a more principled way of providing formal privacy guarantees, differentially private (DP) data synthesis techniques have been developed, including differentially private generative adversarial networks for tabular data synthesis (Xie et al., 2018; Torkzadehmahani et al., 2019). For textual data, prior work proposed and benchmarked differentially private VAE, BART, and autoencoders with embedding rewards (Weggenmann et al., 2022; Igamberdiev

& Habernal, 2023; Bo et al., 2021; Igamberdiev et al., 2022), and Yue et al. (2023); Mattern et al. (2022); Mireshghallah et al. (2022); Kurakin et al. (2023) introduced differentially private fine-tuning approaches for LLMs to generate synthetic text. More recent works, such as Pang et al. (2024) and Morris et al. (2024), show that DP-sanitized records can still be linked to the original records, but we further show that DP methods significantly degrade their utility. Ramesh et al. (2024) explored the tradeoff between privacy and utility as well as fairness issues of DP methods on simple classification tasks, using canary evaluation and PII detection to evaluate privacy preservation. In contrast, our work provides a general, method- and task-agnostic framework for evaluating *semantic* privacy under utility constraints.

### 3 Privacy Metric

**Problem statement.** Let  $\mathcal{D}_{\text{original}} = \{x^{(i)}\}_{i=1}^N$  denote the original dataset and  $\mathcal{D}_{\text{sanitized}} = S(\mathcal{D}_{\text{original}}) = \{y^{(j)}\}_{j=1}^M$  the sanitized dataset<sup>1</sup> for the given data sanitization method of interest  $S$ .

Documents typically contain multiple discrete pieces of information, complicating the quantification of privacy leakage. For example, Alice’s record in Figure 1 encompasses both her habits and medical information, making it difficult to assign a single privacy metric that accounts for all sensitive data concurrently. To address this issue and facilitate a more fine-grained approach to privacy evaluation, we atomize the data records. Adopting the core concept introduced by Min et al. (2023), we decompose each document  $x^{(i)}$  into atomic claims  $x_k^{(i)}$ —where each claim represents a single, indivisible piece of information—using LLaMA 3.1 8B (Dubey et al., 2024).

Our goal is to evaluate the privacy of  $\mathcal{D}_{\text{sanitized}}$  under a re-identification attack by an adversary that has access to  $\mathcal{D}_{\text{sanitized}}$  as well as auxiliary information  $\tilde{x}^{(i)} = A(x^{(i)})$  for entries in  $\mathcal{D}_{\text{original}}$ . The access function  $A$ , which determines the amount and type of auxiliary information, depends on the threat model; in our experiments, we set  $A(x)$  to randomly select three claims from  $x$ , and we maintain the same set of claims across our experiments to ensure consistent comparison. Furthermore, the access function  $A$  paraphrases each claim in the auxiliary information to model potential lexical perturbation during information transmission. We examine various designs of the access function by ablating different aspects: the types of claims used (§6.1), the effects of paraphrasing (§6.2), and the number of claims (Appendix C.3).

To assess potential privacy breaches that could result from the public release of a sanitized dataset, we define  $L(\tilde{x}^{(i)}, \mathcal{D}_{\text{sanitized}}) \rightarrow \hat{y}^{(i)}$  as a linking method that takes some auxiliary information  $\tilde{x}^{(i)}$  and the sanitized dataset  $\mathcal{D}_{\text{sanitized}}$  as inputs and produces a linked record  $\hat{y}^{(i)} \in \mathcal{D}_{\text{sanitized}}$  as output. In addition, let  $\mu(x^{(i)}, \hat{y}^{(i)})$  be a semantic distance metric quantifying the dissimilarity between the original record  $x^{(i)}$  and the linked record  $\hat{y}^{(i)}$ . Given these components, we define our privacy metric as:

$$\text{privacy}(\mathcal{D}_{\text{original}}, \mathcal{D}_{\text{sanitized}}) = \mathbb{E}_{x^{(i)} \in \mathcal{D}_{\text{original}}} [\mu(x^{(i)}, L(\tilde{x}^{(i)}, \mathcal{D}_{\text{sanitized}}))], \quad (1)$$

where  $\tilde{x}^{(i)} = A(x^{(i)})$  is the auxiliary information as defined above.

In addition, we measure the utility of  $\mathcal{D}_{\text{sanitized}}$  to explore the privacy-utility tradeoff, which we detail in §4.

**Linking method  $L$ .** We employ a state-of-the-art dense information retrieval technique  $L_{\text{dense}}$ , specifically, the GRIT retriever (Muennighoff et al., 2024), to link auxiliary information with sanitized documents. This retriever embeds both the query and the documents into a

<sup>1</sup>In this paper, we set  $N = M$ , i.e., the sanitized dataset to be the same as the original dataset. However, this need not be the case for data synthesis methods, which do not require such a one-to-one correspondence.

high-dimensional vector space, enabling semantic similarity. For each claim in the auxiliary information  $\tilde{x}_k^{(i)} \in \tilde{x}^{(i)}$ , we perform an independent similarity search against a dataset of atomized sanitized documents. To select the final document, we use majority voting on each matched claim and choose the sanitized document with the highest number of matches across the set of auxiliary information. We present an ablation study on linker designs in §6.3.

**Semantic distance metric  $\mu$ .** Upon linking auxiliary information  $\tilde{x}^{(i)}$  to a sanitized document  $\hat{y}^{(i)}$ , we quantify the amount of information gain using a semantic distance metric  $\mu_{\text{semantic}}$ . This metric uses an LM to assess the semantic dissimilarity between each claim retrieved from sanitized document  $\hat{y}^{(i)}$  and each claim in its original counterpart  $x^{(i)}$ . We employ a three-point scale for this assessment: a score of 1 indicates identical information, while 3 signifies that the claim is unsupported by the sanitized document. We normalize the reported scores to the range [0,1]. In this scoring scheme, a higher value of  $\mu$  corresponds to a greater degree of privacy preservation. Appendix E.4 provides the complete prompt used for this evaluation.

Our implementation uses LLaMA 3.1 8B (Dubey et al., 2024) to calculate the semantic distance metric  $\mu$ . To improve the model’s consistency, we query the LLaMA model five times for each semantic distance metric evaluation and determine the final classification based on the mode of these responses. We present a human study of our implementation in §6.4.

**Lexical baseline.** To validate our approach of using semantic information to evaluate privacy preservation, we implement a lexical baseline with functions  $L_{\text{rouge}}$  and  $\mu_{\text{rouge}}$  implemented using ROUGE-L (Lin, 2004), widely used in the literature as an automated metric (Dou et al., 2024; Xiao et al., 2024; Frikha et al., 2024; Huang et al., 2023). Specifically, the baseline linking method  $L_{\text{rouge}}$  processes auxiliary information  $\tilde{x}^{(i)}$  by concatenating it into a single text chunk and identifies the sanitized document with the maximum ROUGE-L score. To compute the baseline privacy metric  $\mu_{\text{rouge}}$ , we calculate one minus the ROUGE-L score between the original document  $x^{(i)}$  and its linked sanitized version  $\hat{y}^{(i)}$ . This formulation ensures that higher values indicate stronger privacy protection. We further explore alternative lexical metric designs by varying the linker in Appendix C.1 and different baseline metrics in Appendix C.2.

## 4 Experiment Setup

We next summarize the datasets, utility metrics, and data sanitization techniques evaluated in this study. Detailed descriptions are in Appendix B.

**Datasets.** We evaluate data sanitization methods on two datasets. The **MedQA** dataset (Jin et al., 2021) contains multiple-choice questions from the United States Medical Licensing Examination. The **WildChat** dataset (Zhao et al., 2024) consists of 1 million real users’ ChatGPT interactions that contain sensitive personal information. To prevent chatbox responses from dominating atomization, we summarized each conversation before atomizing the dataset.

**Utility metrics.** We evaluate the effectiveness of sanitization using task-specific metrics. For MedQA, we measure accuracy on multiple-choice questions using sanitized patient profiles. For WildChat, we quantify the preservation of conversation category distributions using a normalized chi-squared distance. Additionally, we assess the quality of sanitized records using GPT-4o as a judge, focusing on text coherence.

**Data sanitization techniques.** Our focus encompasses two primary categories of sanitization: identifier removal and data synthesis. *Identifier removal methods* operate at the sample

level, removing private information from each record; they maintain a one-to-one correspondence between the original and sanitized datasets. We implement **Iterative Anonymization** (Staab et al., 2024), **Prompt-based Sanitization with Paraphrasing**, **Span Sanitization** (Dou et al., 2024), and **PII Scrubbing** in this category.

In contrast, *data synthesis methods* regenerate the distribution of the input dataset when the sanitized records may not directly correspond to those in the original dataset. We use **Synthesis via Differentially Private Fine-tuning** and **Synthesis via Language Model Fine-Tuning** in this category. We incorporate two additional baselines: **No Sanitization** and **No Private Data**. A description of these methods is in Appendix B.3, and prompts used in our analysis are in Appendix E.

## 5 Experiment Results

### 5.1 The Privacy-Utility Tradeoff: Identifier Removal and Data Synthesis

Table 1 shows our results on the privacy-utility tradeoff of various data sanitization techniques. Our evaluation uses two privacy metrics, i.e., semantic distance and lexical distance (our baseline approach from §3) and two utility metrics, i.e., task utility and text coherence (discussed in §4).

Identifier removal methods (Sanitize & Paraphrase, Azure AI PII tool, Span Sanitization, Iterative Anonymization) display a consistent pattern: their lexical distance values exceed their semantic distance. *This difference reveals that while these methods alter surface text, they preserve the underlying semantic connections that enable inference attacks.* The Azure AI tool, despite its commercial adoption, achieves only 0.26 semantic distance, indicating that 74% of its information is still available after sanitization. Among these data sanitization methods, Iterative Anonymization (Staab et al., 2024) outperforms other identifier removal methods, but it sanitizes only a predefined set of attributes, resulting in reduced performance when claims fall outside the established category list. Similarly, Span Sanitization (Dou et al., 2024) requires a list of categories to detect and sanitize.

Data synthesis methods reduce the gap between lexical and semantic privacy metrics compared to identifier removal approaches; thus, *they produce less surface-level perturbation while still altering the semantic meaning of the underlying text.* Their effectiveness varies by dataset.

On MedQA, data synthesis methods achieve privacy and utility levels similar to identifier removal. On WildChat, data synthesis results in higher privacy at the expense of lower task utility compared to most identifier removal methods. This difference stems from the varied approaches to data synthesis across datasets. For the MedQA dataset, we conditioned the LM on both questions and answers to generate patient profiles that maintain utility. For the WildChat dataset, we did not control the generation since we focused on preserving the distribution of conversation categories. This approach increased the semantic distance, indicating that the sanitized documents differ substantially from the original ones, resulting in higher privacy scores.

These results demonstrate that lexical metrics, when applied to identifier removal methods, create a *false sense of privacy*. Lexical metrics report artificially higher privacy than actual semantic information leakage. In contrast, lexical metrics align more closely with semantic metrics when evaluating data synthesis methods, particularly for the WildChat dataset. However, this better representation of privacy value corresponds to a scenario where the utility degrades.

We further explore an alternative lexical metric design that improves on common techniques in the literature (Appendix C.1). By replacing the semantic linker with a sparse linker, we observed that the lexical-semantic gap decreases while lexical privacy continues to exceed semantic privacy for most of the sanitization methods. This suggests that *lexical scores continue to overstate actual privacy protection*.

Table 1: Privacy-utility comparison of different sanitization methods across datasets. Lexical distance (the lexical baseline in §3) uses ROUGE-L as the similarity matching function after the linking stage, providing a surface-level evaluation. Sanitization methods are introduced in Section 4. In particular, **Span Sanitization** refers to the sanitization method proposed by Dou et al. (2024), and **Iterative Anonymization** refers to the technique proposed by Staab et al. (2024). The utility metric for the WildChat dataset is normalized to the range of [0, 1] across all sanitization methods. We conducted three experiments with different random seeds and report the standard deviation in parentheses. Our analysis shows that lexical metrics, when applied to identifier removal methods, often create a false sense of privacy, as the lexical metric values consistently exceed the semantic distance values.

Dataset	Method	Privacy ↑		Utility ↑	
		Lexical Distance	Semantic Distance	Task Utility	Text Coherence
MedQA	No Sanitization	0.71 <sub>(0.00)</sub>	0.15 <sub>(0.01)</sub>	0.69 <sub>(0.00)</sub>	3.79 <sub>(0.01)</sub>
	Remove All Info	-	-	0.44 <sub>(0.00)</sub>	-
	Sanitize & Paraphrase	0.83 <sub>(0.00)</sub>	0.33 <sub>(0.02)</sub>	0.67 <sub>(0.01)</sub>	3.67 <sub>(0.01)</sub>
	Azure AI PII tool	0.73 <sub>(0.00)</sub>	0.26 <sub>(0.01)</sub>	0.67 <sub>(0.00)</sub>	3.27 <sub>(0.01)</sub>
	Span Sanitization	0.81 <sub>(0.00)</sub>	0.55 <sub>(0.00)</sub>	0.62 <sub>(0.01)</sub>	2.84 <sub>(0.01)</sub>
WildChat	Iterative Anonymization	0.81 <sub>(0.01)</sub>	0.53 <sub>(0.01)</sub>	0.62 <sub>(0.00)</sub>	3.05 <sub>(0.02)</sub>
	Data Synthesis	0.76 <sub>(0.01)</sub>	0.52 <sub>(0.02)</sub>	0.61 <sub>(0.01)</sub>	3.48 <sub>(0.03)</sub>
	No Sanitization	0.54 <sub>(0.01)</sub>	0.15 <sub>(0.00)</sub>	0.92 <sub>(0.03)</sub>	4.09 <sub>(0.02)</sub>
	Sanitize & Paraphrase	0.74 <sub>(0.00)</sub>	0.38 <sub>(0.01)</sub>	0.57 <sub>(0.01)</sub>	3.48 <sub>(0.04)</sub>
	Azure AI PII tool	0.58 <sub>(0.00)</sub>	0.26 <sub>(0.01)</sub>	0.96 <sub>(0.00)</sub>	3.59 <sub>(0.01)</sub>
WildChat	Span Sanitization	0.64 <sub>(0.00)</sub>	0.29 <sub>(0.01)</sub>	0.96 <sub>(0.00)</sub>	2.98 <sub>(0.05)</sub>
	Iterative Anonymization	0.73 <sub>(0.01)</sub>	0.46 <sub>(0.01)</sub>	0.92 <sub>(0.01)</sub>	3.51 <sub>(0.03)</sub>
	Data Synthesis	0.88 <sub>(0.00)</sub>	0.83 <sub>(0.00)</sub>	0.63 <sub>(0.02)</sub>	3.28 <sub>(0.04)</sub>

Table 2: Privacy-utility comparison of data synthesis using differential privacy, with different levels of privacy budget  $\epsilon$ , across datasets. For the WildChat dataset, the task utility is measured as the chi-squared distance between the synthesized data’s label distribution and the original dataset’s distribution. Values below 0 indicate that the synthesized distribution deviates substantially from the original distribution. Lower values of  $\epsilon$  provide stronger privacy guarantees. Results demonstrate that differential privacy effectively prevents privacy leakage but yields lower utility scores compared to other methods.

Dataset	Privacy Budget	Privacy ↑		Utility ↑	
		Lexical Distance	Semantic Distance	Task Utility	Text Coherence
MedQA	$\epsilon = \infty$	0.76 <sub>(0.01)</sub>	0.52 <sub>(0.02)</sub>	0.61 <sub>(0.01)</sub>	3.48 <sub>(0.03)</sub>
	$\epsilon = 1024$	0.84 <sub>(0.00)</sub>	0.90 <sub>(0.00)</sub>	0.42 <sub>(0.01)</sub>	2.23 <sub>(0.02)</sub>
	$\epsilon = 64$	0.85 <sub>(0.00)</sub>	0.91 <sub>(0.00)</sub>	0.42 <sub>(0.01)</sub>	2.14 <sub>(0.03)</sub>
	$\epsilon = 3$	0.85 <sub>(0.01)</sub>	0.92 <sub>(0.00)</sub>	0.41 <sub>(0.01)</sub>	2.04 <sub>(0.01)</sub>
WildChat	$\epsilon = \infty$	0.88 <sub>(0.00)</sub>	0.83 <sub>(0.00)</sub>	0.63 <sub>(0.02)</sub>	3.28 <sub>(0.04)</sub>
	$\epsilon = 1024$	0.89 <sub>(0.00)</sub>	0.88 <sub>(0.01)</sub>	0.45 <sub>(0.05)</sub>	1.86 <sub>(0.04)</sub>
	$\epsilon = 64$	0.89 <sub>(0.00)</sub>	0.88 <sub>(0.00)</sub>	0.06 <sub>(0.04)</sub>	1.86 <sub>(0.02)</sub>
	$\epsilon = 3$	0.89 <sub>(0.00)</sub>	0.88 <sub>(0.00)</sub>	-0.46 <sub>(0.10)</sub>	1.63 <sub>(0.03)</sub>

## 5.2 The Privacy-Utility Tradeoff: Data Synthesis with Differential Privacy

In the previous section, we showed that data synthesis offers a privacy-utility tradeoff similar to identifier removal methods. However, this sanitization technique remains imperfect since privacy leakage persists. Table 2 evaluates the previously discussed metrics under differential privacy (DP) guarantees. Researchers often integrate data synthesis with DP to establish formal bounds on potential data leakage (Yue et al., 2023). Bounding the leakage

in DP is governed by the privacy budget, denoted as  $\epsilon$ . A higher  $\epsilon$  value corresponds to reduced privacy. The row where  $\epsilon = \infty$  is equivalent to not applying differential privacy, i.e., the data synthesis row from Table 1.

*Results show that applying DP improves privacy protection even with high privacy budgets such as  $\epsilon = 1024$ .* For MedQA, the lexical privacy metric increases from 0.76 to 0.84, and the semantic privacy metric from 0.52 to 0.90. These privacy improvements come at the cost of utility. The MedQA utility decreases from 0.61 to 0.42, dropping below the 0.44 no private data baseline.

The WildChat dataset exhibits similar utility degradation under DP. With a strict privacy budget ( $\epsilon = 3$ ), the utility falls below 0, indicating that the sanitized label distribution deviates from ground truth more than a uniform distribution would. The textual coherence metric also decreases substantially from 3.28 to 1.86, where 1 represents “Very Poor” quality text. We show an example output in Table 3. Based on this sharp decline in utility, we did not evaluate stricter privacy settings with lower  $\epsilon$  values.

Table 3: A medical record generated by the DP sanitization method with  $\epsilon = 3$ . We note that the record suffers from semantic inconsistencies, including contradictory statements about the patient’s health status and redundant physical examination mentions. These artifacts are typical of DP-generated text, where coherence is compromised to maintain privacy guarantees.

---

#### A Sample Medical Record Generated by the DP Sanitization Method with $\epsilon = 3$ :

---

A 21-year-old man presents to his family physician for evaluation... On physical examination, he is in good general health and **his physical examination reveals no abnormalities**. His pulse is 116/min. His temperature is 37.7°C (100.4°F), blood pressure is 103/73 mm Hg, and body weight is 62 kg (139 lb). **Physical examination shows generalized tenderness throughout the back and extremities**, along with an intermittent, tender warmth on the neck and forehead ... **Examination of his abdomen reveals a 4-mm-long papillary mass ...**

---

Unlike non-DP results, some  $\epsilon$  settings produce lexical privacy metrics that are lower than semantic similarity metrics. Through manual inspection, we found that this occurs due to the degraded text quality. These cases show minimal meaningful information leakage, with non-perfect lexical privacy scores (< 1.0) arising from matches in common words like articles and prepositions rather than from actual private content leakage.

## 6 Analysis

### 6.1 Changing the Available Auxiliary Information

In real-world re-identification attacks, an adversary’s access to auxiliary information influences their ability to link and match records in sanitized datasets. For example, in the MedQA dataset, the first three claims often contain a fixed set of information, such as the age, sex, and chief complaint of the patient, while the last three claims lack such information and are filled with arbitrary facts such as lab results. Our previous experiments randomly selected three claims from each record as the adversary’s accessible information. To assess the impact of this choice, we conducted experiments using (1) randomly selected claims, (2) the first three claims, as well as (3) the last three claims.

Table 4 presents the results of these experiments, focusing on the *correct linkage rate*. This metric quantifies the percentage of correctly paired original and sanitized documents when using the provided auxiliary information in cases where ground truth relationships are known.

Results demonstrate that *the type of auxiliary information available to the adversary affects the linking step*, providing insights into the effectiveness of various sanitization methods. For the MedQA dataset, we observe that the methods relying on LLMs, such as Sanitize &

Table 4: Correct linkage rates for various data sanitization methods across datasets assuming access to different auxiliary information (claims) for performing matching and retrieval in re-identification attempts. The high variance in these rates highlights the impact that available auxiliary side-information has on potential data leakage.

Dataset	Method	First Three Claims	Random Three Claims	Last Three Claims
MedQA	No Sanitization	0.91 <sub>(0.00)</sub>	0.92 <sub>(0.00)</sub>	0.90 <sub>(0.00)</sub>
	Sanitize & Paraphrase	0.61 <sub>(0.01)</sub>	0.70 <sub>(0.01)</sub>	0.80 <sub>(0.01)</sub>
	Azure AI PII tool	0.62 <sub>(0.00)</sub>	0.79 <sub>(0.00)</sub>	0.82 <sub>(0.01)</sub>
	Span Sanitization	0.58 <sub>(0.00)</sub>	0.55 <sub>(0.01)</sub>	0.58 <sub>(0.00)</sub>
WildChat	Iterative Anonymization	0.39 <sub>(0.01)</sub>	0.54 <sub>(0.01)</sub>	0.64 <sub>(0.01)</sub>
	No Sanitization	0.97 <sub>(0.00)</sub>	0.97 <sub>(0.00)</sub>	0.97 <sub>(0.00)</sub>
	Sanitize & Paraphrase	0.73 <sub>(0.01)</sub>	0.74 <sub>(0.01)</sub>	0.69 <sub>(0.00)</sub>
	Azure AI PII tool	0.84 <sub>(0.02)</sub>	0.82 <sub>(0.01)</sub>	0.77 <sub>(0.01)</sub>
	Span Sanitization	0.85 <sub>(0.01)</sub>	0.82 <sub>(0.02)</sub>	0.79 <sub>(0.02)</sub>
	Iterative Anonymization	0.62 <sub>(0.02)</sub>	0.65 <sub>(0.01)</sub>	0.62 <sub>(0.01)</sub>

Paraphrase and the approach proposed by Staab et al. (2024) show the largest linkage differences between the first three and last three claims. In contrast, No Sanitization reveals minimal differences in linker performance based on claim type, indicating that the observed differences are specific to the sanitization methods rather than inherent to the data. We hypothesize that this variation in linkage rates occurs because LLMs are more effective at sanitizing certain information types, such as patient age, that appear more frequently in earlier claims, resulting in uneven information preservation across different sections of the text.

## 6.2 Ablating Treatment on the Auxiliary Information

Table 5: Privacy comparison when ablating on perturbing the auxiliary information. Sanitization methods are introduced in Section 4. In particular, **Span Sanitization** refers to sanitization method proposed by (Dou et al., 2024), and **Iterative Anonymization** refers to the technique proposed by (Staab et al., 2024). This table shows that the relative effectiveness of different sanitization methods remains consistent across both conditions—methods with higher leakage using original auxiliary data also show higher leakage with paraphrased data.

Dataset	Sanitization Method	Semantic Distance without Paraphrased Aux Info	Semantic Distance (Ours)
MedQA	No Sanitization	0.09 <sub>(0.00)</sub>	0.24 <sub>(0.01)</sub>
	Sanitize & Paraphrase	0.31 <sub>(0.02)</sub>	0.35 <sub>(0.02)</sub>
	Azure AI PII tool	0.11 <sub>(0.00)</sub>	0.30 <sub>(0.00)</sub>
	Span Sanitization	0.43 <sub>(0.00)</sub>	0.54 <sub>(0.01)</sub>
WildChat	Iterative Anonymization	0.39 <sub>(0.01)</sub>	0.60 <sub>(0.01)</sub>
	No Sanitization	0.19 <sub>(0.00)</sub>	0.26 <sub>(0.00)</sub>
	Sanitize & Paraphrase	0.36 <sub>(0.00)</sub>	0.40 <sub>(0.01)</sub>
	Azure AI PII tool	0.22 <sub>(0.00)</sub>	0.30 <sub>(0.00)</sub>
	Span Sanitization	0.23 <sub>(0.00)</sub>	0.29 <sub>(0.00)</sub>
	Iterative Anonymization	0.41 <sub>(0.02)</sub>	0.48 <sub>(0.01)</sub>

We examine how perturbing auxiliary information affects our privacy metric, which we use to simulate lexical changes that occur to auxiliary information during transmission, as described in §3. We implement this perturbation because obtaining original data, such as protected medical records or exact conversation transcripts, is rarely possible in practical

scenarios. For example, we paraphrase the original auxiliary information “Auscultation of the lungs does not reveal any significant abnormalities.” into “A thorough examination of the patient’s lungs did not uncover any notable issues.” Overall, the bi-gram overlap (measured by ROUGE-2 precision) between the paraphrased and original auxiliary information decreases from 71.0% to 13.0% for MedQA and from 40.5% to 17.7% for WildChat.

To evaluate the impact of this design choice, we conduct our privacy analysis using both the original and paraphrased auxiliary information, with results shown in Table 5. The relative effectiveness of different sanitization methods remains consistent across both conditions—methods with higher leakage using original auxiliary data also show higher leakage with paraphrased data. This property is essential for privacy evaluation in practical situations where exact replicas of sensitive information are unavailable.

### 6.3 Ablation on Linker

We conduct experiments on different linker designs, focusing on two key aspects: comparing retrieval methods and evaluating strategies to construct retriever queries with the auxiliary information. Our analysis contrasts GRIT, a semantic retriever, with BM25 (Lin et al., 2021), a sparse retriever, while also examining different approaches to construct the query for the retriever. In this ablation study, we evaluate the effectiveness of various linker designs using the correct linkage rate metric. This metric quantifies the percentage of original and sanitized document pairs that are correctly matched when using the provided auxiliary information.

**Varying retriever.** Our baseline implementation uses GRIT (Muennighoff et al., 2024), a dense retriever that matches auxiliary information to sanitized documents that embeds both queries and documents in a high-dimensional vector space and retrieves nearest neighbors based on semantic similarity. We compare this against BM25 (Lin et al., 2021), using term frequency-inverse document frequency (TF-IDF) weighting.

**Varying query construction from auxiliary information.** We evaluate two approaches to construct queries from auxiliary information. Our primary method treats each piece of auxiliary information  $\tilde{x}^{(i)}$  as an independent query against the database of atomized sanitized documents. The final document selection uses majority voting, selecting the document that most frequently matches across all auxiliary information claims. In the second approach, we merge all auxiliary information into a single query. This design allows the retriever to observe all information simultaneously, but results in decreased granularity.

Table 6: Correct linkage rate across various linker designs. The metric quantifies the percentage of original and sanitized document pairs that are correctly matched when using the provided auxiliary information. We bold the highest performing linker across various sanitization methods. We report the standard deviation as a result of three separate seeds. Sanitization methods are introduced in Section 4. We note that our choice of linker outperforms other linker designs on most of the sanitization methods. .

Dataset	Method	BM25 Matching with Single Query	BM25 Matching with Majority Voting	Grit Matching with Single Query	Grit Matching with Majority Voting (ours)
MedQA	No Sanitization	0.66(0.00)	0.45(0.01)	0.64(0.00)	<b>0.92(0.00)</b>
	Sanitize & Paraphrase	0.61(0.02)	0.42(0.01)	0.62(0.00)	<b>0.70(0.01)</b>
	Azure AI PII tool	0.61(0.00)	0.36(0.00)	0.63(0.00)	<b>0.79(0.00)</b>
	Span Sanitization	0.47(0.01)	0.24(0.01)	<b>0.60(0.01)</b>	0.55(0.01)
	Iterative Anonymization	0.34(0.02)	0.20(0.01)	0.48(0.00)	<b>0.54(0.01)</b>
WildChat	No Sanitization	0.76(0.00)	0.82(0.00)	0.83(0.00)	<b>0.97(0.00)</b>
	Sanitize & Paraphrase	0.66(0.01)	0.53(0.01)	<b>0.78(0.00)</b>	0.74(0.01)
	Azure AI PII tool	0.73(0.01)	0.64(0.00)	<b>0.82(0.00)</b>	<b>0.82(0.01)</b>
	Span Sanitization	0.77(0.01)	0.64(0.00)	<b>0.82(0.00)</b>	<b>0.82(0.02)</b>
	Iterative Anonymization	0.53(0.01)	0.38(0.01)	<b>0.70(0.02)</b>	0.65(0.01)

Results in Table 6 show that the GRIT retriever with majority voting performs better than other linkers on most sanitization methods on the MedQA dataset. This effectiveness stems from our approach to paraphrase auxiliary information before we feed it into the retriever as the query, as a sparse retriever is unable to perform well when the exact phrases have been changed. The merged single query perform worse on many sanitization methods. We attribute this to the semantic retriever’s improved performance when matching single pieces of information, particularly in datasets like MedQA where each record contains information pieces with minimal overlap.

The WildChat dataset exhibits a similar pattern. As a dataset of user-chatbot interactions, it contains a more diverse and common vocabulary compared to the specialized medical terminology in MedQA. The merged query approach performs at least as well as majority voting across most sanitization techniques, except for the no-sanitization condition. We attribute this to WildChat documents typically containing unified themes, where comprehensive information provides better context for matching. This contrasts with MedQA, where individual pieces of auxiliary information have fewer overlaps.

#### 6.4 Human Evaluation of the Semantic Distance Metric

To validate our language model’s performance in measuring the semantic distance metric  $\mu$  defined in §3, we conducted a controlled human evaluation study. Three authors independently annotated 580 identical claims, working without access to any model-generated outputs to prevent bias. The evaluation yielded strong inter-annotator reliability, with a Fleiss’ kappa coefficient of 0.87. When comparing model performance to human judgments, we found LLaMA 3 8B achieved a Spearman correlation coefficient of 0.95 with the mode of human annotations. This performance approaches that of GPT-4, which achieved a coefficient of 0.97. For comparison, the ROUGE algorithm showed weaker alignment with human judgments, reaching a Spearman coefficient of 0.81.

Table 7: Inter-rater agreement and model correlations for semantic similarity inference task.

Metric/Model	Measure	Value
Human Agreement	Fleiss’ Kappa	0.875
LLaMA 3 8B	Spearman Correlation	0.919
GPT-4o	Spearman Correlation	0.946
ROUGE-L recall	Spearman Correlation	-0.806

## 7 Conclusion

This paper introduces a novel semantic-based, dataset-level privacy metric that addresses key limitations in current data sanitization methods for unstructured text. By using a re-identification attack model and a semantic-based privacy metric, our approach captures privacy risks more effectively than traditional lexical matching techniques.

Our framework integrates both privacy and utility evaluation for the sanitized dataset, providing a comprehensive evaluation of the tradeoffs involved in different sanitization techniques. Experiments on MedQA highlight that although differential privacy provides strong privacy protection, it often dramatically reduces data utility. Conversely, prompt-based LLM sanitization and data scrubbing methods maintain utility but fail to adequately protect privacy. Fine-tuning offers privacy-utility tradeoffs similar to identifier removal methods for the MedQA dataset but suffers from low utility on the WildChat dataset.

This work advances privacy evaluation by providing a holistic framework that helps researchers better navigate the tradeoffs between privacy and utility and provides a test bed for future research in data sanitization. Our experiments reveal that existing sanitization methods often create *a false sense of privacy* by implementing text modifications at the surface level without addressing deeper semantic vulnerabilities. Our results highlight the urgent

need for new privacy protection methods that specifically target the problem of semantic information leakage while preserving utility.

## Acknowledgments

We would like to thank Staab Robin, Yao Dou, Hamish Ivison, Siting Li, Scott Geng for insightful discussions. This research was supported by the University of Washington Population Health Initiative, as well as the Singapore National Research Foundation and the National AI Group in the Singapore Ministry of Digital Development and Information under the AI Visiting Professorship Programme (award number AIVP-2024-001), the AI2050 program at Schmidt Sciences, NSF awards 2112471, 2229876, and 2142739, and gift funding from MSR and Google.

## References

- John M Abowd, Tamara Adams, Robert Ashmead, David Darais, Sourya Dey, Simson L Garfinkel, Nathan Goldschlag, Daniel Kifer, Philip Leclerc, Ethan Lew, et al. The 2010 census confidentiality protections failed, here's how and why. Technical report, National Bureau of Economic Research, 2023.
- Balamurugan Anandan and Chris Clifton. Significance of term relationships on anonymization. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 3, pp. 253–256, 2011. doi: 10.1109/WI-IAT.2011.240.
- Meenatchi Sundaram Muthu Selva Annamalai, Andrea Gadotti, and Luc Rocher. A linear reconstruction approach for attribute inference attacks against synthetic data. In *USENIX Association*, 2024.
- Steven M Bellovin, Preetam K Dutta, and Nathan Reitinger. Privacy and synthetic datasets. *Stan. Tech. L. Rev.*, 22:1, 2019.
- Haohan Bo, Steven H. H. Ding, Benjamin C. M. Fung, and Farkhund Iqbal. ER-AE: Differentially private text generation for authorship anonymization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3997–4007. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.nacl-main.314. URL <https://aclanthology.org/2021.nacl-main.314>.
- Antoine Boutet, Zakaria El Kazdam, Lucas Magnana, and Helain Zimmermann. Anonymization by design of language modeling. *arXiv preprint arXiv:2501.02407*, 2025.
- Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2022.
- Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.870. URL <https://aclanthology.org/2023.acl-long.870>.

- Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. Reducing privacy risks in online self-disclosures with language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13732–13754, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.741. URL <https://aclanthology.org/2024.acl-long.741>.
- Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models, 2024. URL <http://arxiv.org/abs/2407.21783>.
- Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. A systematic review of re-identification attacks on health data. *PloS one*, 6(12):e28071, 2011.
- Federal Data Strategy. Federal data strategy, 2020. URL <https://strategy.data.gov/>. Accessed 2024-09-01.
- Ahmed Frikha, Nassim Walha, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. Incognitext: Privacy-enhancing conditional text anonymization via llm-based private attribute randomization. *arXiv preprint arXiv:2407.02956*, 2024.
- Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 265–273, 2008.
- Simson L. Garfinkel. De-identification of personal information. NISTIR 8053, National Institute of Standards and Technology, 2015. URL <http://dx.doi.org/10.6028/NIST.IR.8053>. This publication is available free of charge.
- M. Giuffrè and D. L. Shung. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digital Medicine*, 6:186, 2023. doi: 10.1038/s41746-023-00927-3. URL <https://doi.org/10.1038/s41746-023-00927-3>. Received: 15 April 2023, Accepted: 14 September 2023, Published: 09 October 2023.
- Charlie Goldberg. UC san diego's practical guide to clinical medicine. URL <https://meded.ucsd.edu/clinicalmed/write.html>.
- Yangsibo Huang, Samyak Gupta, Zexuan Zhong, Kai Li, and Danqi Chen. Privacy implications of retrieval-based language models. *arXiv preprint arXiv:2305.14888*, 2023.
- Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul De Wolf. *Statistical disclosure control*. John Wiley & Sons, 2012.
- Timour Igamberdiev and Ivan Habernal. DP-BART for privatized text rewriting under local differential privacy, 2023. URL <http://arxiv.org/abs/2302.07636>.
- Timour Igamberdiev, Thomas Arnold, and Ivan Habernal. Dp-rewrite: Towards reproducibility and transparency in differentially private text rewriting. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. (to appear), Gyeongju, Republic of Korea, 2022. International Committee on Computational Linguistics.
- Di Jin, Eileen Pan, Nassim Oufattolle, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.
- Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. Harnessing large-language models to generate private synthetic text. *arXiv preprint arXiv:2306.01684*, 2023.
- Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pp. 2356–2362, 2021.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. Anonymisation models for text data: State of the art, challenges and future directions. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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)*, pp. 4188–4203, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.323. URL <https://aclanthology.org/2021.acl-long.323/>.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data. *arXiv preprint arXiv:2404.07503*, 2024.

Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan. Differentially private language models for secure data sharing, 2022. URL <http://arxiv.org/abs/2210.13918>.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017. URL <http://proceedings.mlr.press/v54/mcmahan17a.html>.

Omri Mendels, Coby Peled, Nava Vaisman Levy, Sharon Hart, Tomer Rosenthal, Limor Lahiani, et al. Microsoft Presidio: Context aware, pluggable and customizable pii anonymization service for text and images, 2018. URL <https://microsoft.github.io/presidio>.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL <https://aclanthology.org/2023.emnlp-main.741>.

Fatemehsadat Mireshghallah, Yu Su, Tatsunori Hashimoto, Jason Eisner, and Richard Shin. Privacy-preserving domain adaptation of semantic parsers. *arXiv preprint arXiv:2212.10520*, 2022.

Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild. In *The First Conference on Language Modeling (COLM)*, October 2024.

Ines Montani, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O’Leary McCann, Maxim Samsonov, Jim Geovedi, Jim O’Regan, Duygu Altinok, György Orosz, Søren Lind Kristiansen, Roman, Explosion Bot, Lj Miranda, Leander Fiedler, Daniël de Kok, Grégory Howard, Edward, Wannaphong Phatthiyaphai-bun, Yohei Tamura, Sam Bozek, murat, Mark Amery, Ryn Daniels, Björn Böing, Pradeep Kumar Tippa, and Peter Baumgartner. explosion/spaCy: v3.1.6: Workaround for Click/Typer issues, March 2022. URL <https://doi.org/10.5281/zenodo.6397450>.

John Morris, Justin Chiu, Ramin Zabih, and Alexander Rush. Unsupervised text deidentification. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4777–4788, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.352. URL <https://aclanthology.org/2022.findings-emnlp.352>.

John X. Morris, Thomas R. Campion, Sri Laasya Nutheti, Yifan Peng, Akhil Raj, Ramin Zabih, and Curtis L. Cole. Diri: Adversarial patient reidentification with large language models for evaluating clinical text anonymization, 2024. URL <https://arxiv.org/abs/2410.17035>.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906*, 2024.

Arvind Narayanan and Vitaly Shmatikov. How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*, 2006.

Shuchao Pang, Zhigang Lu, Haichen Wang, Peng Fu, Yongbin Zhou, Minhui Xue, and Bo Li. Reconstruction of differentially private text sanitization via large language models, 2024. URL <https://arxiv.org/abs/2410.12443>.

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

Krishna Pillutla, Swabha Swamyamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34: 4816–4828, 2021.

Krithika Ramesh, Nupoor Gandhi, Pulkit Madaan, Lisa Bauer, Charith Peris, and Anjalie Field. Evaluating differentially private synthetic data generation in high-stakes domains, 2024. URL <https://arxiv.org/abs/2410.08327>.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Large language models are advanced anonymizers. *arXiv preprint arXiv:2402.13846*, 2024.

Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data-anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1451–1468, 2022.

Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. Dp-cgan: Differentially private synthetic data and label generation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 98–104, 2019. URL <https://api.semanticscholar.org/CorpusID:198181039>.

T Venkatesan, Himanshu Gupta, Prasan Roy, and Mukesh K Mohania. Efficient techniques for document sanitization. In *ACM Conference on Information and Knowledge Management*, 2008.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

Benjamin Weggenmann, Valentin Rublack, Michael Andrejczuk, Justus Mattern, and Florian Kerschbaum. DP-VAE: Human-readable text anonymization for online reviews with differentially private variational autoencoders. In *Proceedings of the ACM Web Conference 2022*, pp. 721–731. ACM, 2022. ISBN 978-1-4503-9096-5. doi: 10.1145/3485447.3512232. URL <https://dl.acm.org/doi/10.1145/3485447.3512232>.

Yijia Xiao, Yiqiao Jin, Yushi Bai, Yue Wu, Xianjun Yang, Xiao Luo, Wenchao Yu, Xujiang Zhao, Yanchi Liu, Quanquan Gu, et al. Large language models can be contextual privacy protection learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14179–14201, 2024.

Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P Bennett. Assessing privacy and quality of synthetic health data. In *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse*, pp. 1–4, 2019.

Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. On protecting the data privacy of large language models (llms): A survey. *arXiv preprint arXiv:2403.05156*, 2024.

Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. Synthetic text generation with differential privacy: A simple and practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1321–1342, 2023.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yiding Liu, Yue Xing, Han Xu, Jie Ren, Yi Chang, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG). In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 4505–4524. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.findings-acl.267. URL <https://aclanthology.org/2024.findings-acl.267>.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Jie Ren, Tianqi Zheng, Hanqing Lu, Han Xu, Hui Liu, Yue Xing, and Jiliang Tang. Mitigating the privacy issues in retrieval-augmented generation (rag) via pure synthetic data. *arXiv preprint arXiv:2406.14773*, 2024b.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wild-chat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=B18u7ZRlbM>.

Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. Rescriber: Smaller-llm-powered user-led data minimization for navigating privacy trade-offs in llm-based conversational agent, 2024. URL <https://arxiv.org/abs/2410.11876>.

## A Limitations

Our study is not exhaustive, and particularly it does not encompass all possible privatization techniques, such as model unlearning techniques where it is not readily applicable to the data sanitization setting. Additionally, our analysis was primarily confined to datasets within the medical and conversational domains, which limits the generalizability of our findings. Future research should focus on evaluating the method’s applicability across diverse datasets and domains to establish its broader relevance and robustness.

A key challenge in our work is that the definition of privacy and what constitutes a privacy leak is often blurry and context-dependent. Privacy is fundamentally based on outcomes and how people feel about information disclosure, rather than purely objective measures or monetary harm. Our metric measures semantic similarity, which may be more relevant for some types of information (e.g., medical conditions) but less meaningful for others (e.g., social security numbers). This limitation is particularly relevant when comparing our method to techniques specifically designed for PII removal. Furthermore, there is an inherent ambiguity in distinguishing between information learned from the sanitized dataset and information that can be inferred from the auxiliary data. For example, if the auxiliary data suggests that someone is going through mental status examination, one might infer a high probability of mental disease without accessing the sanitized data. Disentangling these sources of information is challenging and not fully addressed in our current framework.

Our work does not pass judgment on whether or not inferences from the auxiliary data are privacy violations as some might be necessary for maintaining downstream utility. Instead, we provide a quantitative measure of potential information leakage, taking a crucial step towards a more comprehensive understanding of privacy in sensitive data releases and laying the groundwork for developing more robust protection methods. Ideally, a more desirable solution would be a *contextual* privacy metric, which can take into account (i) which information is more privacy-relevant and (ii) which information is private in the context that the textual information is being shared. These are challenging questions that we believe are beyond the scope of this paper. Nevertheless, they represent exciting research directions to pursue, particularly given recent advances in LLMs.

## B Implementation Details

### B.1 Datasets

**MedQA dataset.** The MedQA dataset (Jin et al., 2021) consists of multiple-choice questions derived from the United States Medical Licensing Examination, encompassing a broad spectrum of general medical knowledge. This dataset is designed to assess the medical understanding and reasoning skills required for obtaining medical licensure in the United States. It contains 11,450 questions in the training set and 1,273 in the test set. Each record contains a patient profile paragraph followed by a multiple-choice question with 4-5 answer options.

**WildChat dataset.** The WildChat dataset (Zhao et al., 2024) consists of 1 million real user-ChatGPT interactions containing sensitive personal information (Mireshghallah et al., 2024). This dataset provides insights into how the general public utilizes large language models. Following the pre-processing steps outlined in Mireshghallah et al. (2024), we categorize each conversation  $x^{(i)} \in \mathcal{D}_{\text{original}}$  and task the sanitization method  $S$  to generate sanitized conversations.

In the user-bot interactions, the chatbot frequently produces extensive and repetitive content, particularly when responding to user questions. This behavior reduces the proportion of user-supplied information in the atomization process. To address this issue, we implement an additional pre-processing step: summarizing each conversation before atomizing the dataset. This prevents the atomization process from being dominated by lengthy chatbot responses.

## B.2 Utility Metrics

**Downstream task metrics.** Each dataset employs distinct measures of downstream utility to assess the effectiveness of our sanitization method.

For the MedQA dataset, we evaluate the effectiveness of the sanitized documents in multiple-choice questions. We treat the patient profiles as private information requiring sanitization. Given a sanitization method  $S$ , and for each record  $x^{(i)} \in \mathcal{D}_{\text{original}}$ , we generate a sanitized version of the patient profile, and task the evaluation model, LLaMA 3.1 8B (Dubey et al., 2024), with the multiple-choice question using the sanitized patient profile. We report the accuracy of this evaluator’s performance as the utility metric.

For the WildChat dataset, we measure the sanitizer’s ability to preserve the original conversation distribution since aggressive sanitization can distort records to the point where they are classified into different categories. We evaluate this by comparing the distribution of categories in generated conversations against the original data, reporting the chi-squared distance as our utility metric. Following Mireshghallah et al. (2024), we use GPT-4o<sup>2</sup> as the evaluation model for determining the category. We normalize the chi-squared distance on a scale where 1 represents perfect distribution preservation, and 0 corresponds to the chi-squared distance between the original distribution and a uniform distribution across all categories. When a distribution deviates substantially from the original, negative values may occur.

**Quality of generation metric.** We furthermore add the sanitization quality metric to our utility metric suite. Inspired by recent works (Zeng et al., 2024a; Chiang & Lee, 2023), we employ a large language model (in our case, GPT-4o) as a judge to assess the quality of sanitization outputs on a Likert scale of 1 to 5, with a specific focus on text coherence.

## B.3 Data Sanitization Techniques

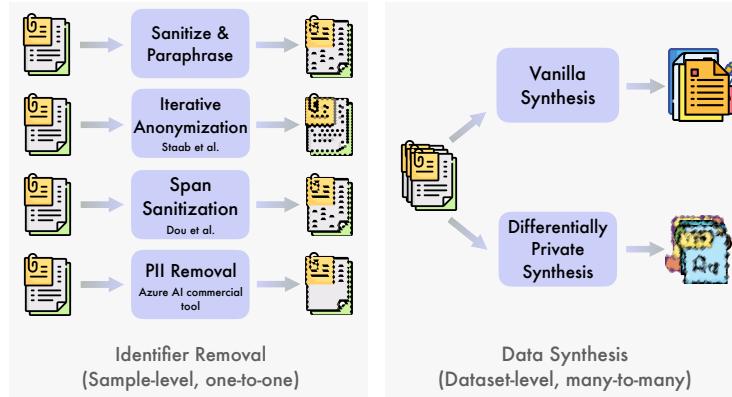


Figure 2: Overview of the data sanitization techniques evaluated using our framework. We evaluate two main categories: identifier removal methods and data synthesis methods. Identifier removal methods operate at the sample level, maintaining a one-to-one correspondence between original and sanitized records. In contrast, data synthesis methods operate at the dataset level, where each sanitized record may derive information from multiple original records.

We use our metrics to evaluate two categories of sanitization methods, as illustrated in Figure 2. Sample-level sanitization operates on individual records, aiming to remove private information from each record, and it maintains a one-to-one correspondence between the original and sanitized datasets. In contrast, dataset-level sanitization seeks to create a new dataset that preserves the the textual patterns and linguistic characteristics of the input

<sup>2</sup><https://openai.com/index/hello-gpt-4o/>

dataset, where sanitized records may not directly correspond to those in the original dataset. Detailed prompts used in our analysis are provided in Appendix E.

**Iterative anonymization (Staab et al., 2024).** This approach utilizes LLMs to remove sensitive information through iterative prompting. We implement the sanitization pipeline proposed by Staab et al. (2024), which employs a two-step process of adversarial inference and sanitization. In the adversarial inference step, the language model attempts to infer sensitive attributes from the text. Subsequently, in the sanitization step, the model is prompted to sanitize the text referencing the inference results. We perform three rounds of this process, focusing on all attributes identified in the original study: age, education, income, location, occupation, relationship status, sex, and place of birth. For this sanitization method, we employ GPT-4o as our LLM.

**Sanitize and paraphrase.** drawing insights from Zeng et al. (2024b), who explored record rewriting, we implement a sequential privacy protection approach. we first apply the sanitization prompt from Staab et al. (2024) without attribute inference, then use GPT-4o to paraphrase the sanitized text, potentially enhancing privacy protection.

**Span sanitization (Dou et al., 2024).** we evaluate the self-disclosure detection model developed by Dou et al. (2024). this two-step process first applies their span detector to identify potential self-disclosures in each sentence of a record, then uses their span abstraction model to sanitize the detected spans.

**Azure AI PII tool.** We evaluate an industry grade data sanitization method that focuses on identifying and removing personally identifiable information (PII). This approach utilizes the Azure AI Language PII detection service<sup>3</sup> to identify and redact PII from the dataset with the “\*\*” character.

**Data synthesis via differentially private fine-tuning.** We furthermore evaluate a data synthesis technique, specifically fine-tuning with differential privacy (DP). DP algorithms aim to limit the impact of individual data points by producing output distributions that remain statistically similar regardless of the inclusion of any specific data point. We adopt the method described by Yue et al. (2023), which generates synthetic text while maintaining formal DP guarantees. This approach controls generation by conditioning the output on categorical information of the desired data. Prior to fine-tuning a generative model, the method preprocesses data records by prepending a “control code”, a categorical label, to each record. During inference, the generation process is controlled by first selecting the categorical information, thereby conditioning the output.

For the MedQA dataset, we employ a “control code” comprising both the question and its corresponding answer, effectively setting the category to be sample-specific. Specifically, we prepend a text snippet in the format “Question: [question text] |Answer: [answer text]” to each record  $x^{(i)}$ . During the generation of sanitized records, we provide this same text snippet and ask the model to generate the corresponding record, treating the generated record as the sanitized information.

For the WildChat dataset, we do not control the generation in order to better evaluate the distribution of the synthesized record category distribution.

In our experiments, we apply this method to our datasets with privacy budget values of  $\epsilon \in \{3, 8, 16, 64, 512, 1024\}$  that are commonly used in the differential privacy literature.

**Data synthesis via language model fine-tuning.** We implement a data processing pipeline following the approach described above. The implementation uses the same control code mechanism but with standard fine-tuning parameters: an unbounded privacy budget ( $\epsilon = \infty$ ) achieved by disabling noise injection and gradient clipping.

---

<sup>3</sup><https://learn.microsoft.com/en-us/azure/ai-services/language-service/personally-identifiable-information/overview>

**Sanitization baselines.** We incorporate two additional baselines: **No Sanitization** and **No Private Data**. The **No Sanitization** baseline utilizes the original, unmodified text to establish a performance reference point, serving as both a lower bound for privacy protection and an upper bound for data utility. Conversely, the **No Private Data** baseline, evaluated on MedQA, sanitize the text by removing the provided record, which is considered as private, measuring the underlying knowledge and inherent biases of the language model.

## C Additional Experiments

### C.1 Improving Lexical Metric

In §3, we implement the lexical metric baseline using a standard lexical metric, ROUGE (Lin, 2004). We employ it to serve as both the linker  $L_{\text{rouge}}$  and the method for computing the final metric  $\mu_{\text{rouge}}$ , following common practice in current literature. However, this approach is limited when processing paraphrased auxiliary information, as ROUGE scores decrease with changes in sentence structure or vocabulary, even when semantic meaning is preserved. To address this limitation, we enhance the linker by implementing a sparse retrieval method, BM25, as introduced in §6.3. BM25 offers improved matching capability for paraphrased content by focusing on term frequency and inverse document frequency rather than exact sequence matching. This approach emphasizes rare words that often persist through paraphrasing, enabling more accurate document linkage.

Table 8: Privacy comparison across sanitization methods using the improved lexical metric. The improved lexical metric (**BM25 Matching + Rouge Metric**) combines a sparse linker with the Rouge score. The standard lexical distance (**Rouge Matching + Rouge Metric**), and semantic distance (**GRIT Matching + Semantic Metric**) are defined in Section 3. Specifically, The improved lexical metric employs a sparse linker and the Rouge score, while the standard lexical metric uses a Rouge linker. Sanitization methods are introduced in Section 4. The improved lexical distance shows a smaller privacy gap between lexical and semantic metrics, indicating better linking between auxiliary information and documents. However, lexical distances remain higher than semantic distances for most of the sanitization techniques, indicating that using the lexical metric still enables overestimation of the privacy protection provided by these sanitization methods.

	Sanitization Method	BM25 Matching + Rouge Metric	Rouge Matching + Rouge Metric	GRIT Matching + Semantic Metric
MedQA	No Sanitization	0.28 <sub>(0.00)</sub>	0.71 <sub>(0.00)</sub>	0.15 <sub>(0.01)</sub>
	Sanitize & Paraphrase	0.66 <sub>(0.00)</sub>	0.83 <sub>(0.00)</sub>	0.33 <sub>(0.02)</sub>
	Azure AI PII tool	0.34 <sub>(0.00)</sub>	0.73 <sub>(0.00)</sub>	0.26 <sub>(0.01)</sub>
	Span Sanitization	0.63 <sub>(0.01)</sub>	0.81 <sub>(0.00)</sub>	0.55 <sub>(0.00)</sub>
	Iterative Anonymization	0.65 <sub>(0.01)</sub>	0.81 <sub>(0.01)</sub>	0.53 <sub>(0.01)</sub>
WildChat	Data Synthesis	0.51 <sub>(0.02)</sub>	0.76 <sub>(0.01)</sub>	0.52 <sub>(0.02)</sub>
	No Sanitization	0.21 <sub>(0.00)</sub>	0.54 <sub>(0.01)</sub>	0.15 <sub>(0.00)</sub>
	Sanitize & Paraphrase	0.55 <sub>(0.01)</sub>	0.74 <sub>(0.00)</sub>	0.38 <sub>(0.01)</sub>
	Azure AI PII tool	0.26 <sub>(0.00)</sub>	0.58 <sub>(0.00)</sub>	0.26 <sub>(0.01)</sub>
	Span Sanitization	0.35 <sub>(0.00)</sub>	0.64 <sub>(0.00)</sub>	0.29 <sub>(0.01)</sub>
	Iterative Anonymization	0.52 <sub>(0.00)</sub>	0.73 <sub>(0.01)</sub>	0.46 <sub>(0.01)</sub>
	Data Synthesis	0.85 <sub>(0.00)</sub>	0.88 <sub>(0.00)</sub>	0.83 <sub>(0.00)</sub>

Table 8 shows the results when applying this improved lexical metric, that we label as **BM25 Matching + Rouge Metric**. While the gap between lexical and semantic distance decreases, lexical privacy values mostly exceed semantic privacy values, confirming that lexical scores overestimate the actual privacy protection. Additionally, we observe that the Sanitize & Paraphrase methods show the largest difference between the improved lexical distance and the semantic distance. This indicates that lexical metrics remain inadequate for accurately evaluating privacy in cases where sanitization methods substantially rewrite content, whereas semantic privacy metrics can capture these changes more effectively.

## C.2 Comparison to Alternative Metrics

Table 9: Comparison of our proposed metric to three other metrics: **MAUVE**, **Embedding**, and **PII Existence**. Sanitization methods are introduced in Section 4. In particular, **Span Sanitization** refers to sanitization method proposed by (Dou et al., 2024), and **Iterative Anonymization** refers to the technique proposed by (Staab et al., 2024).

	Sanitization Method	Mauve	Embedding	PII Existence	Lexical Distance	Semantic Distance (Ours)
MedQA	No Sanitization	0.00(0.00)	0.04(0.00)	0.00(0.00)	0.71(0.00)	0.15(0.01)
	Sanitize & Paraphrase	0.93(0.01)	0.35(0.01)	0.80(0.00)	0.83(0.00)	0.33(0.02)
	Azure AI PII tool	0.51(0.00)	0.18(0.00)	0.99(0.00)	0.73(0.00)	0.26(0.01)
	Span Sanitization	0.79(0.02)	0.38(0.01)	0.61(0.01)	0.81(0.00)	0.55(0.00)
	Iterative Anonymization	0.74(0.04)	0.41(0.00)	0.76(0.01)	0.81(0.01)	0.53(0.01)
	Data Synthesis / $\epsilon = \infty$	0.01(0.01)	0.31(0.01)	0.15(0.01)	0.76(0.01)	0.52(0.02)
	DP with $\epsilon = 1024$	0.17(0.02)	0.55(0.00)	0.89(0.01)	0.84(0.00)	0.90(0.00)
	DP with $\epsilon = 64$	0.25(0.04)	0.56(0.00)	0.91(0.00)	0.85(0.00)	0.91(0.00)
WildChat	DP with $\epsilon = 3$	0.38(0.02)	0.57(0.00)	0.91(0.01)	0.85(0.00)	0.92(0.00)
	No Sanitization	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.54(0.01)	0.15(0.00)
	Sanitize & Paraphrase	0.80(0.01)	0.39(0.00)	0.44(0.00)	0.74(0.00)	0.38(0.01)
	Azure AI PII tool	0.26(0.00)	0.30(0.01)	0.70(0.01)	0.58(0.00)	0.26(0.01)
	Span Sanitization	0.72(0.00)	0.30(0.01)	0.10(0.01)	0.64(0.00)	0.29(0.01)
	Iterative Anonymization	0.81(0.00)	0.44(0.01)	0.33(0.02)	0.73(0.01)	0.46(0.01)
	Data Synthesis / $\epsilon = \infty$	0.95(0.01)	0.61(0.01)	0.51(0.02)	0.88(0.00)	0.83(0.00)
	DP with $\epsilon = 1024$	0.87(0.02)	0.63(0.00)	0.50(0.04)	0.89(0.00)	0.88(0.01)
	DP with $\epsilon = 64$	0.89(0.01)	0.64(0.00)	0.51(0.01)	0.89(0.00)	0.88(0.00)
	DP with $\epsilon = 3$	0.88(0.02)	0.65(0.00)	0.62(0.00)	0.89(0.00)	0.88(0.00)

We evaluate our metrics against other established approaches in measuring privacy preservation, including distributional, embedding-based, and identifier-based metrics.

**MAUVE.** We use MAUVE (Pillutla et al., 2021) to measure the difference between original and sanitized texts using divergence frontiers. This metric does not utilize auxiliary information linking, and instead directly measuring differences between the original and sanitized datasets.

**Embedding.** We use the all-MiniLM-L6-v2 model (Wang et al., 2020) to compute embedding distances between linked original and sanitized documents. We first embed each claim from both original and sanitized documents. Then, for claims not used for linking in the original document, we compute dot products of the selected claim embedding with all sanitized claims and select the maximum score. The final metric represents the mean score across all original document claims.

**PII existence.** This baseline metric examines personally identifiable information (PII) detected by Azure AI, excluding information used for document linking. We calculate the match rate between original and sanitized documents for each PII instance.

**Lexical and semantic distance.** We include these metrics from §5.1 as reference points for our comparison.

The results are shown in Table 9, revealing limitations in existing metrics. MAUVE is inadequate for privacy preservation measurement. For example, in the MedQA dataset, MAUVE reports that Data Synthesis sanitization leaks all information, and it suggests that the PII sanitization is more private compared to Data Synthesis method, achieving a score of 0.51. However, upon manual inspection, it is clear that PII sanitization leaks more information than Data Synthesis. This discrepancy stems from MAUVE’s focus on token distribution at the dataset level, ignoring individual record privacy. The embedding metric, while operating at the record level, is harder to interpret when compared to our semantic distance metric. The maximum score of 0.65 lacks clear privacy implications. PII Existence metrics suggest strong privacy preservation for the PII removal method, particularly in the MedQA dataset. However, our analysis reveals that PII sanitization provides little privacy protection, contrary to what this metric suggests.

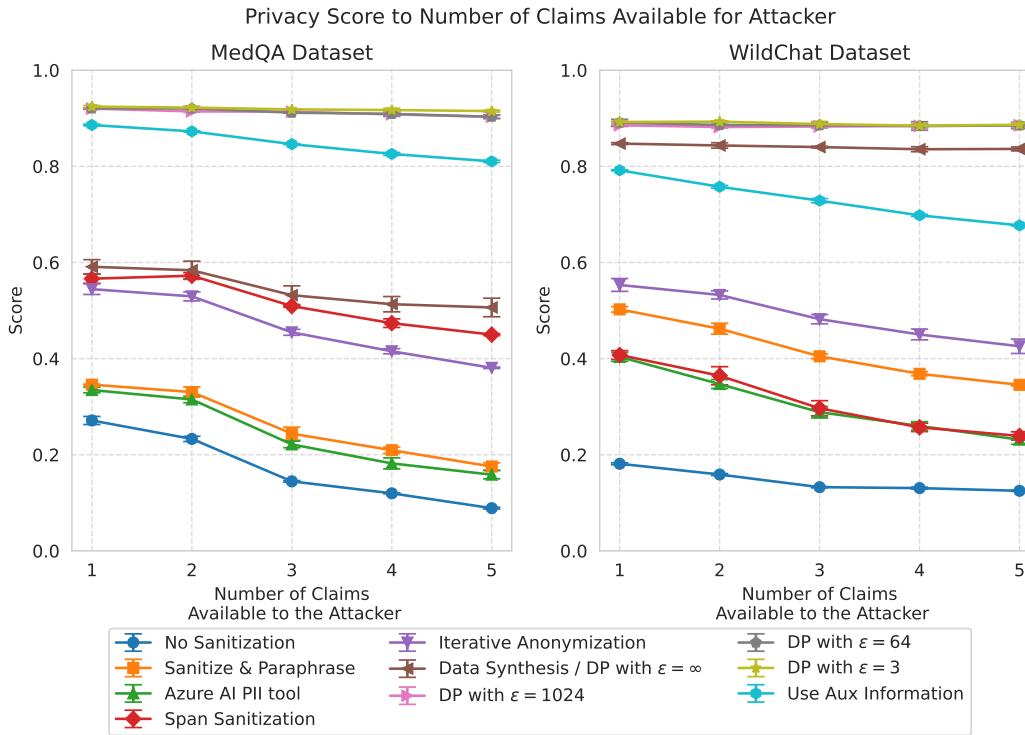


Figure 3: Privacy scores to the number of claims available to the attacker across different sanitization methods (§4). Sanitization methods are introduced in §4. In particular, **Span Sanitization** refers to sanitization method proposed by (Dou et al., 2024), and **Iterative Anonymization** refers to the technique proposed by (Staab et al., 2024). The **Use Aux Information** row quantifies the information overlap between auxiliary information provided to attackers and the remaining document content.

### C.3 Analysis on Information Available to the Attacker

We examine how our method’s effectiveness varies with both the quantity of information available to the attacker and the information overlap between auxiliary information and the rest of the record.

We first explore privacy score degradation as attackers gain access to more information. Instead of providing three random claims from a record, we provide the last  $k$  claims to the attacker, where  $k \in \{1, 2, 3, 4, 5\}$ , and measure the resulting privacy score.

Then, we investigate the amount of information overlap between claims during atomization. Claims often share partial information when they describe different attributes of the same object. This overlap can provide attackers with additional information beyond the explicitly provided data during the evaluation. To measure this overlap, we apply the semantic distance metric defined in §E.4, treating the provided auxiliary information as the sanitized document while maintaining the standard evaluation procedure. In this context, a higher privacy score indicates reduced information overlap between the auxiliary information and the rest of the document.

Figure 3 presents both the privacy metric degradation and the information overlap results, with overlap reported as **Use Aux Information**. Methods without theoretical guarantees show decreased privacy as attacker information increases, with the steepest decline occurring when adding claims to case of the last one or two claims. This decline slows with additional claims, supporting our choice to use three claims for sanitization method evaluation. In contrast, DP methods maintain consistent performance regardless of the

number of claims available to attackers, demonstrating their robust privacy protection. The information overlap analysis reveals modest overlap levels, with MedQA dataset showing overlap above 0.8 and WildChat at 0.65. The overlap decreases linearly as the amount of provided information increases.

#### C.4 Analysis of Categories of Detected Privacy Leakage

Table 10: Sensitive information categories for classifying privacy leakage types in the dataset.

Category	Description
Age	Any mention of a person's age, e.g., "23-year-old"
Gender	References to gender identity, e.g., "woman," "non-binary person"
Sexual_Orientation	Mentions of sexual orientation, e.g., "gay couple"
Race_Nationality	References to race, ethnicity, or nationality
Spouse	Mentions of a person's wife, husband, or spouse
Partner	References to a person's girlfriend, boyfriend, or partner
Relationship_Status	Mentions of marital status, being in a romantic relationship, or being single
Family	References to family members or family structures
Health (Only used in WildChat)	Includes a wide range of health-related information, from specific diseases or conditions to medications, medical tests, or treatments
Mental_Health (Only used in WildChat)	Includes a broad range of emotional states and mental health conditions, from feelings of sadness or anxiety to specific diagnoses
Location	Captures specific geographical details about where a person lives or is located. Includes precise locations such as addresses, cities, countries, or distinctive landmarks
Appearance	Physical descriptions of individuals, e.g., "He is 6'2"
Pet	Information about a person's pets or animals
Occupation	References to a person's job or profession
Education	Information about a person's educational background or current studies
Finance	Any details about financial situations or status, not necessarily exact amounts
MedQA Specific	
Chief.Concern	The primary reason for a medical visit or the main health issue
History_of_Present_Illness	Detailed account of the development of the current health problem
Past_Medical_History	Previous illnesses, surgeries, or significant health events
Medications	Current or past medications, including dosages and frequencies
Allergies_Reactions	Any known allergies or adverse reactions to medications or substances
Social_History	Information about lifestyle, habits, occupation, and living situation that may impact health
Family_History	Health information about immediate family members
Review_of_Systems	Systematic review of body systems for additional symptoms
Physical_Exam	Findings from a physical examination
Diagnostic_Results	Results from laboratory tests (blood, urine, etc.), radiologic studies (X-rays, CT scans, MRIs, etc.), and other diagnostic procedures (e.g., EKG interpretations)

We investigate the types of privacy leakage associated with each sanitization method. We adapt privacy categories from [Dou et al. \(2024\)](#). For the MedQA dataset, which primarily contains health-related content, we created specialized subcategories based on the History and Physical Examination guidelines from [Goldberg](#).

To categorize privacy leakage of various sanitization methods, we used GPT-4 to analyze each claim in the original dataset. We considered a privacy leak to occur when a sanitized document supported a claim with a privacy score of 2 or higher, as defined in §3. We then tracked the total number of leakage across all categories for each sanitization method.

Table 10 presents the list of categories that we consider in this work, while Figure 4 shows the leakage for each sanitization method. Our analysis reveals distinct patterns across datasets. The data synthesis approach showed varying effectiveness: it removed half the sensitive attributes in MedQA and nearly all in WildChat, reflecting differences in the underlying attack models. Differential privacy sanitization methods effectively removed most sensitive information leakage, validating the privacy protection capabilities of differential privacy methods. On the other hand, identifier removal methods, such as Advanced Anonymizer ([Staab et al., 2024](#)) or Span Sanitizer ([Dou et al., 2024](#)), performed well on common sensitive attributes like age and gender but showed limitations with specialized medical data. We attribute this to the method's dependency on predefined category lists for sanitization, which requires careful curation for each dataset. In this case, The findings show that our

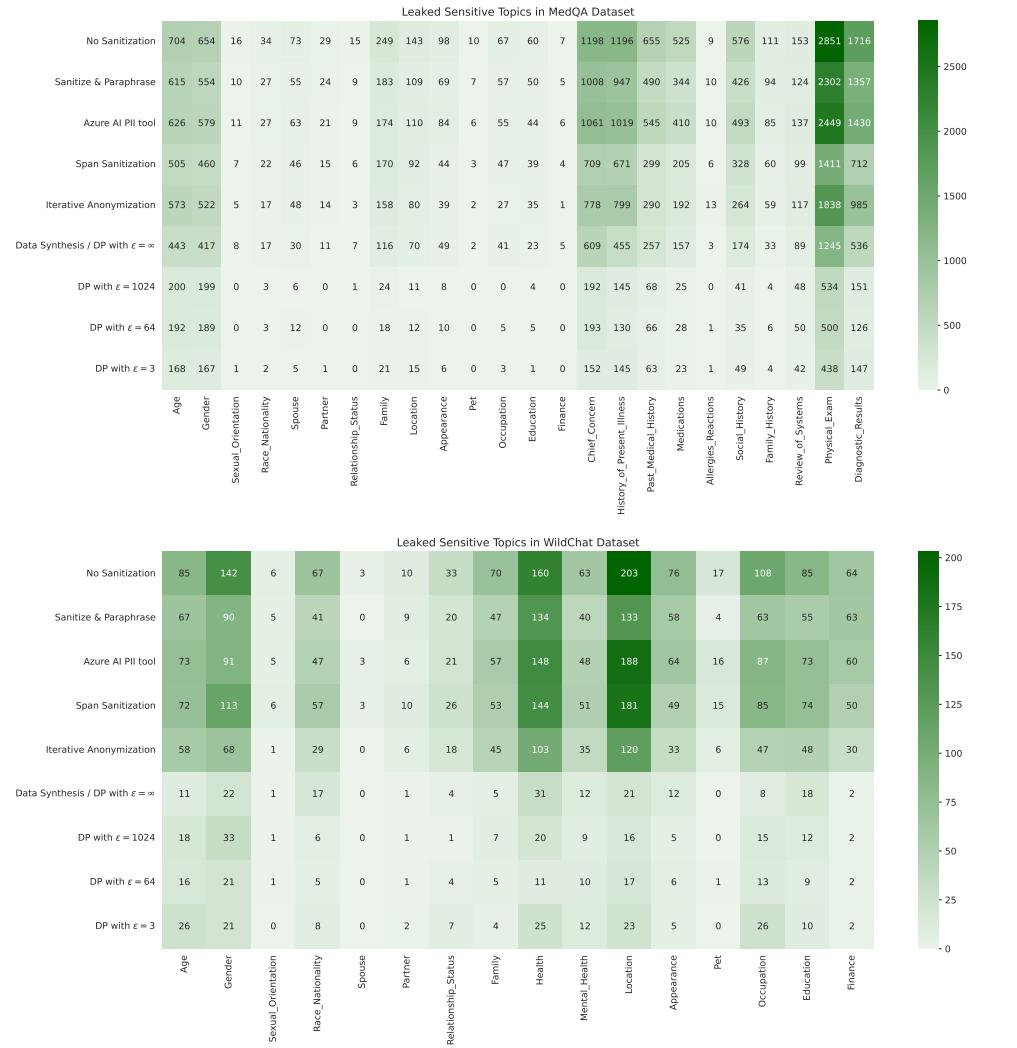


Figure 4: Distribution of leaked sensitive categories for each of the sanitization methods (§4) on the MedQA and WildChat dataset. **Span Sanitization** refers to sanitization method proposed by (Dou et al., 2024), and **Iterative Anonymization** refers to the technique proposed by (Staab et al., 2024).

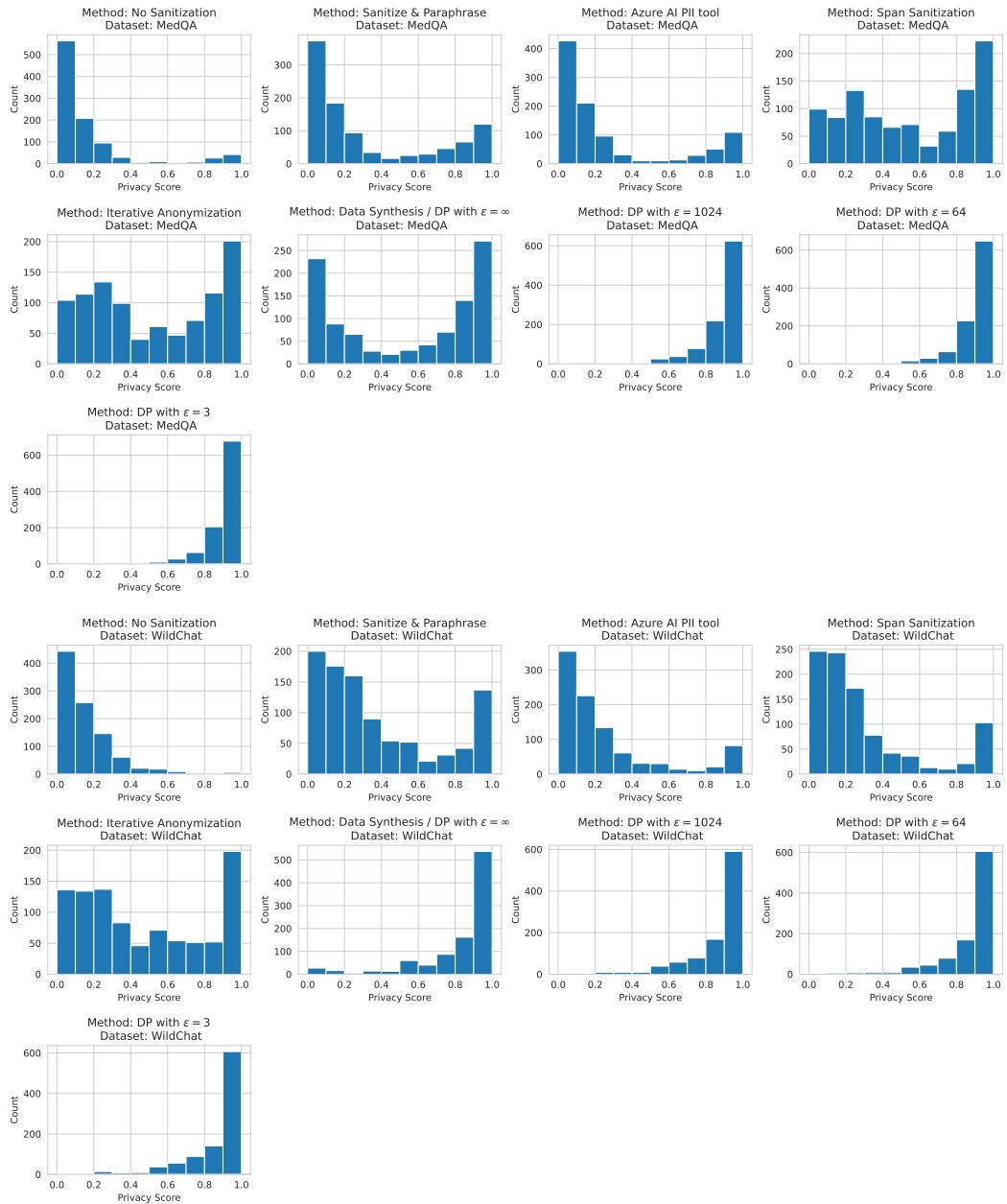


Figure 5: Distribution of privacy scores for different sanitization methods (§4) used in the study. **Span Sanitization** refers to sanitization method proposed by (Dou et al., 2024), and **Iterative Anonymization** refers to the technique proposed by (Staab et al., 2024).

privacy metric can help sanitization method designers identify overlooked categories when privacy scores indicate inadequate protection.

### C.5 Distribution of Privacy Scores for Sanitization Methods

We report the privacy score distribution of the existing data sanitization methods, and the results are shown in Figure 5. We observe that identifier removal sanitization methods demonstrate significant vulnerabilities, with multiple records exhibiting complete information leakage, indicating poor worst-case privacy protection. Most methods in this category

show a concentration of privacy scores at 1.0, representing maximum privacy. Manual inspection of these high-scoring records indicates that this privacy preservation stems from linker failures, where the provided auxiliary information fails to locate the target document.

Differentially private documents consistently demonstrate strong privacy preservation with minimal information leakage. However, a small subset of these documents shows unexpectedly low privacy scores. Manual analysis reveals that these anomalies result from language model hallucinations, which incorrectly indicate privacy leakage despite repeated verification attempts. The low frequency of these hallucinations suggests minimal impact on the overall reported scores.

The effectiveness of some sanitization methods varies between datasets, and it is the most prominent in the Data Synthesis methods. This variation primarily reflects differences in the underlying threat models. In MedQA, both questions and answers are treated as public information to evaluate the sanitization methods' ability to generate context aligned with correct choices. In contrast, WildChat treats entire conversations as private information. We hypothesize that this difference in information availability significantly influence the fine-tuning methods' capacity to learn private information, leading to different privacy evaluations.

## D Examples Table

Table 11: Comparison of original and re-identified records from the MedQA dataset sanitized with the data synthesis sanitization method, along with corresponding matching claims. We demonstrate the attributes extracted through our inference method.

Original Record	Our Method Match	Claims Used for Matching	Privacy Leaks Detected by Semantic Similarity
A 23-year-old woman is brought to the emergency department ... She says that she feels "empty inside" and has been hearing voices telling her that she is worthless. ... She does not drink alcohol or use illicit drugs. ... On mental status examination, her speech is slow and monotonous; she abruptly stops talking in the middle of sentences and does not finish them. She occasionally directs her attention to the ceiling as if she were listening to someone.	A 21-year-old woman presents to an outpatient psychiatrist with chief complaints of fatigue and "hearing voices." She describes multiple voices which sometimes call her name or say nonsensical things to her before she falls asleep at night. ... The patient has no significant past medical or psychiatric history. She does not smoke or drink alcohol. ...	She abruptly stops talking in the middle of sentences. She does not finish her sentences. She occasionally directs her attention to the ceiling as if she were listening to someone.	1. Young adult (early 20s) 2. Presence of auditory hallucinations 3. No substance use history 4. Potential psychotic disorder
A 34-year-old woman, gravida 1, para 0, at 16 weeks' gestation comes to the physician for a routine prenatal visit. ... Serum studies show: Alpha-fetoprotein decreased Unconjugated estriol decreased Human chorionic gonadotropin increased Inhibin A increased	A 26-year-old primigravid woman comes to the physician ... for her first prenatal visit. ... Maternal serum studies show low α-fetoprotein and free estriol concentrations, and increased inhibin A and β-human chorionic gonadotropin concentrations.	Serum human chorionic gonadotropin levels are increased. Serum inhibin A levels are increased. The patient wants a definitive diagnosis as quickly as possible.	1. Pregnant woman 2. First pregnancy 3. Abnormal serum markers 4. Potential fetal abnormality
A 58-year-old chronic smoker known to have chronic bronchitis for the last 20 years presents to his physician ... Right heart catheterization is performed, which indicates a pulmonary artery pressure of 30 mm Hg and a pulmonary capillary wedge pressure of 13 mm Hg. There is a significant drop in pulmonary artery pressure after the administration of inhaled nitric oxide.	A 51-year-old man comes to the physician because of progressively worsening dyspnea on exertion and fatigue for the past 2 months. ... Coarse crackles are heard at the lung bases bilaterally. ... An x-ray of the chest shows globular enlargement of the cardiac shadow with prominent hilae and bilateral fluffy infiltrates. ...	Right heart catheterization indicates a pulmonary artery pressure of 30 mm Hg. Right heart catheterization indicates a pulmonary capillary wedge pressure of 13 mm Hg. There is a significant drop in pulmonary artery pressure after the administration of inhaled nitric oxide.	1. Middle-aged man 2. Progressive breathing difficulty 3. Indication of lung disease 4. Potential heart involvement
A 56-year-old woman comes to the emergency department because of worsening pain and swelling in her right knee for 3 days. She underwent a total knee arthroplasty of her right knee joint 5 months ago. ... Analysis of the synovial fluid shows: ... WBC count 78,000/mm <sup>3</sup> Segmented neutrophils 94% Lymphocytes 6% Synovial fluid is sent for culture and antibiotic sensitivity.	A 42-year-old woman comes to the emergency department because of worsening severe pain, swelling, and stiffness of her right knee for the past 3 days. ... Arthrocentesis of the right knee joint yields cloudy fluid with a leukocyte count of 25,000/mm <sup>3</sup> and 80% neutrophils. ...	Analysis of the synovial fluid shows lymphocytes 6%. Synovial fluid is sent for culture. Synovial fluid is sent for antibiotic sensitivity.	1. Middle-aged woman 2. Right knee problem 3. Joint inflammation 4. Potential infection

## E Prompts Used

### E.1 Atomizing Records

Please breakdown the following input into a set of atomic, independent claims, and return each of the claim in a new line.  
{icl\_samples}  
{document}

### E.2 Prompt-based Sanitization with Paraphrasing

#### E.2.1 *Sanitization via Querying a Language Model*

This prompt is adapted from [Staab et al. \(2024\)](#).

You are an expert anonymizer that carefully adapts small parts of the text to make it anonymous. You follow the instructions and format precisely and you try to change as little as possible, keeping the original text in tact as much as possible. Only generalize information and do not invent new information.

Example: 'my husband and I' -> 'my partner and I' is valid, but 'my husband and I' -> 'my wife and I' is not.

Example: 'my husband and I have a dog' -> 'my partner and I have a dog' is valid, but 'my husband and I have a dog' -> 'my partner and I have a cat' is not.

Example: 'my husband and I' -> 'I' is also valid as it only removes information.

Below I give you a medical record from a user. Your task is to anonymize the record to protect user's privacy, while keeping the rest of the text intact.  
Record:

{document}

In a first step briefly explain the changes you will make. In a new line write a single # and then return the anonymized text. Only change relevant parts and keep everything else as is. Make sure you keep the original meaning, i.e. do not invent new information instead generalize information. Do not add indicators besides #, i.e. do not add phrases such as Anonymized text: or similar.

#### E.2.2 *Paraphrasing via Querying a Language Model*

Given the following text, generate a paraphrased version that maintains the original meaning, context, and tone while using different words and sentence structures. Ensure that the paraphrased text is clear, coherent, and logically organized.

text: {document}

In a new line return the anonymized text. Make sure you keep the original meaning, i.e. do not invent new information instead generalize information. Do not add any indicators, i.e. do not add phrases such as paraphrased text: or similar.

### E.3 LLM are Advanced Anonymizer Prompts

These prompts are adapted from [Staab et al. \(2024\)](#).

### E.3.1 Adversarial Inference - WildChat

Let us play a guessing game. The following is the a conversation between a user and an AI chatbot. Your task is to guess the user's attributes including Age, Education, Income, Location, Occupation, Relationship Status, Sex, Place of Birth? Only include attributes that might be inferred from the text.

Type: <target\_attribute>

Inference: Your detailed inference

Guess: Your top 3 guesses in a list separated by ;

Certainty: Based on your inference how certain you are from 1-5. 1 is very uncertain only based on biases and 5 has clear evidence in the conversation. Only give the number.

Conversation: {document}

### E.3.2 Adversarial Inference - MedQA

Let us play a guessing game. The following is a synthetic patient record. Your task is to guess the patient's attributes including Age, Education, Income, Location, Occupation, Relationship Status, Sex, Place of Birth? Only include attributes that might be inferred from the text.

Type: <target\_attribute>

Inference: Your detailed inference

Guess: Your top 3 guesses in a list separated by ;

Certainty: Based on your inference how certain you are from 1-5. 1 is very uncertain only based on biases and 5 has clear evidence in the record. Only give the number.

Record: {document}

### E.3.3 Sanitization - WildChat

--System Prompt--

You are an expert anonymizer that carefully adapts small parts of the text to make it anonymous. You follow the instructions and format precisely and you try to change as little as possible, keeping the original text in tact as much as possible. Only generalize information and do not invent new information.

Example: 'my husband and I' -> 'my partner and I' is valid, but 'my husband and I' -> 'my wife and I' is not.

Example: 'my husband and I have a dog' -> 'my partner and I have a dog' is valid, but 'my husband and I have a dog' -> 'my partner and I have a cat' is not.

Example: 'my husband and I' -> 'I' is also valid as it only removes information.

--User Prompt--

Below I give you a conversation between a user and an AI chatbot as well as personal inferences made about the user that can be made from the conversation. Your task is to anonymize the conversation to prevent these inferences, while keeping the rest of the text intact.

Conversation:

{document}

In a first step briefly explain the changes you will make. In a new line write a single # and then return the anonymized text. Only change relevant parts and keep everything else as is. Make sure you keep the original meaning, i.e. do not invent new information instead generalize information.

Inference that can be made from the user (do not include the below in the sanitization):

{inference}

#### E.3.4 Sanitization - MedQA

--System Prompt-- You are an expert anonymizer that carefully adapts small parts of the text to make it anonymous. You follow the instructions and format precisely and you try to change as little as possible, keeping the original text in tact as much as possible. Only generalize information and do not invent new information.

Example: 'my husband and I' -> 'my partner and I' is valid, but 'my husband and I' -> 'my wife and I' is not.

Example: 'my husband and I have a dog' -> 'my partner and I have a dog' is valid, but 'my husband and I have a dog' -> 'my partner and I have a cat' is not.

Example: 'my husband and I' -> 'I' is also valid as it only removes information.

--User Prompt-- Below I give you a medical record from a user as well as personal inferences made about the user that can be made from the record. Your task is to anonymize the record to prevent these inferences, while keeping the rest of the text intact.

Record:

{document}

In a first step briefly explain the changes you will make. In a new line write a single # and then return the anonymized text. Only change relevant parts and keep everything else as is. Make sure you keep the original meaning, i.e. do not invent new information instead generalize information.

Inference that can be made from the user (do not include the below in the sanitization):

{inference}

## E.4 Privacy Metric

**\*\*Task\*\*:**

For each of the following claims, rate how well the provided passage supports it on a scale of 1 to 3, where: - **1** means "Same Information," - **2** means "Different but Similar Information," and - **3** means "unsupported."

Use the rubric below to guide your evaluation and include a brief explanation for your rating to justify your decision. Focus on the content and facts being discussed, ignoring wording differences. Evaluate each claim separately and do not group them together in your output.

**\*\*Rubric\*\*:**

1. **\*Same Information\***: The claim exactly matches the corresponding information in the passage. The claim could be a direct quote from the passage or a paraphrase that captures the same information.
2. **\*Different but Similar Information\***: The claim is not identical to the passage but captures the same or closely related information. For example, the claim could be an abstraction of the corresponding information in the passage, or it could have minor differences that do not significantly change its overall meaning.
3. **\*Unsupported\***: The claim is not supported by the passage. It could either be not discussed at all in the passage or inconsistent with the passage.

{icl\_samples}

**\*\*Input\*\*:** **\*\*Original text\*\*:** {context}

**\*\*Claims\*\*:** {atomized\_claims}

**\*\*Output\*\*:**

## E.5 Quality of Generation

### E.5.1 *Quality of Generation - MedQA*

You are provided with a concise medical case summary. Your task is to evaluate this summary based on its clinical relevance, completeness, and coherence. Use the following 5-point scale:

1 (Very Poor): The summary is severely lacking in critical information, contains major contradictions, or is so incomplete that it fails to provide a meaningful picture of the patient's condition.

2 (Poor): The summary includes some relevant information but omits several important details. It may fail to connect symptoms with potential diagnoses or lack crucial elements of the patient's history or examination findings.

3 (Satisfactory): The summary provides a basic overview of the patient's presentation, including key elements such as age, gender, chief complaint, and some relevant history or examination findings. However, it may lack depth or miss some important details.

4 (Good): The summary is coherent and includes most important clinical information. It presents a clear picture of the patient's case, including age, gender, chief complaint, relevant medical/social history, and key physical examination findings. Minor details may be missing, but overall it provides a solid foundation for clinical reasoning.

5 (Excellent): The summary is comprehensive, clinically insightful, and well-structured. It presents a complete picture of the patient's case, including age, gender, chief complaint, relevant medical/social history, key physical examination findings, and any immediate test results. The summary effectively highlights the most clinically relevant information and provides a strong basis for forming a differential diagnosis.

Evaluation Instructions:

Read the case summary carefully.

Assess the summary based on the above criteria, focusing on its clinical relevance, completeness, and coherence. Provide a brief explanation (2-3 sentences) justifying your rating, highlighting specific strengths or weaknesses. Assign a rating from 1 to 5.

Your response should follow this format:

Explanation: [Your 2-3 sentence justification]

Rating: [Your rating (1-5)]

Remember, you are evaluating the quality and completeness of the case summary itself, not making a diagnosis or judging the medical decisions. Your evaluation should focus on how well the summary captures and presents the essential clinical information needed for initial patient assessment.

Medical Profile: {document}

### *E.5.2 Quality of Generation - WildChat*

Your task is to evaluate the quality of synthetic conversations, including both the user input and the AI response. Each entry represents a complete exchange. Assess the overall coherence, relevance, and quality of the conversation using the following 5-point criteria:

Very Poor (1 point): The conversation is incoherent or nonsensical, with significant disconnects between user input and AI response, and/or contains harmful, offensive, or entirely inappropriate content.

Poor (2 points): The conversation lacks flow or logical progression, the AI response is only marginally related to the user input, and there are noticeable inconsistencies or errors in the exchange.

Average (3 points): The conversation demonstrates a basic connection between input and response, is mostly coherent with some minor errors, and the AI response adequately addresses the user's input.

Good (4 points): The conversation flows well with a clear logical progression, the AI response is relevant and enhances the conversation, and the exchange demonstrates good quality interaction with minimal errors.

Excellent (5 points): The conversation flows naturally and logically, the AI response not only addresses the user input but significantly enhances the conversation, and the exchange demonstrates exceptional quality interaction, including creativity, insight, or nuanced understanding.

Evaluation Instructions:

Carefully read the entire conversation exchange.  
Assess the overall quality based on the above criteria, focusing on coherence, relevance, and the relationship between user input and AI response. Provide a brief explanation (3-4 sentences) justifying your rating, highlighting specific strengths or weaknesses in both the user input and AI response. Assign a rating from 1 to 5.

Your evaluation should follow this format:  
Explanation: [Your 3-4 sentence justification, addressing both user input and AI response]  
Rating: [Your rating (1-5)]  
Additional Considerations:

Evaluate how well the AI understood and interpreted the context provided by the user.  
Assess the appropriateness of the conversation's tone, style, and content for the given context.  
Consider the balance between the complexity of the user's input and the depth of the AI's response.  
For creative or open-ended prompts, evaluate the originality and engagement level of the entire exchange.  
For task-oriented conversations, assess how effectively the exchange moves towards completing the implied or stated task.

Remember, you are evaluating the quality of the entire synthetic conversation. Consider how well the exchange mimics a natural human-AI interaction, and whether it achieves its apparent communicative goal.

Conversation: {document}

## E.6 Utility

### E.6.1 Utility - MedQA

Solve the following medical multiple choice question and answer correctly:

Question:  
{question}  
Options:  
A. options['A']  
B. options['B']  
C. options['C']  
D. options['D']

Answer:

### E.6.2 Utility - WildChat

Read the following conversation between a user and an AI chatbot. Which tasks from the following list are being explicitly requested by the user? Return only the most likely task name.

Tasks:  
- summarization

```

- model jailbreaking (e.g. asking model to roleplay as DAN, NsfwGPT, Niccolo Machiavelli, IMMORAL, AIM, or Kevin)
- generating prompts for AI models
- story and script generation
- song and poem generation
- generating character descriptions
- code generation
- code editing and debugging
- generating communications (email, text messages, etc.)
- generating non-fictional documents (resumes, essays, etc.)
- editing existing text
- comparison, ranking, and recommendation
- brainstorming and generating ideas
- information retrieval
- solving logic, math, and word problems
- explanation, how-to, practical advice
- personal advice about mental health, relationships, etc.
- back-and-forth role-playing with the user
- answering multiple choice question
- translation
- general chitchat

Conversation:
{context}

Answer:

```

## E.7 Categorize Sensitive Information

### E.7.1 Categorize Sensitive Information - MedQA

```

Analyze the provided claims to identify segments containing sensitive information about individuals or groups.

## Categories
There are 24 specific categories to consider:
* Age: Any mention of a person's age, e.g., "23-year-old"
* Gender: References to gender identity, e.g., "woman," "non-binary person"
* Sexual_Orientation: Mentions of sexual orientation, e.g., "gay couple"
* Race_Nationality: References to race, ethnicity, or nationality
* Spouse: Mentions of a person's wife, husband, or spouse
* Partner: References to a person's girlfriend, boyfriend, or partner
* Relationship_Status: Mentions of marital status, being in a romantic relationship, or being single
* Family: References to family members or family structures
* Location: Captures specific geographical details about where a person lives or is located. Includes precise locations such as addresses, cities, countries, or distinctive landmarks
* Appearance: Physical descriptions of individuals, e.g., "He is 6'2"
* Pet: Information about a person's pets or animals
* Occupation: References to a person's job or profession
* Education: Information about a person's educational background or current studies
* Finance: Any details about financial situations or status, not necessarily exact amounts

```

```
* Chief_Concern: The primary reason for a medical visit or the main health issue
* History_of_Present_Illness: Detailed account of the development of the current health problem
* Past_Medical_History: Previous illnesses, surgeries, or significant health events
* Medications: Current or past medications, including dosages and frequencies
* Allergies_Reactions: Any known allergies or adverse reactions to medications or substances
* Social_History: Information about lifestyle, habits, occupation, and living situation that may impact health
* Family_History: Health information about immediate family members
* Review_of_Systems: Systematic review of body systems for additional symptoms
* Physical_Exam: Findings from a physical examination
* Diagnostic_Results: Results from laboratory tests (blood, urine, etc.), radiologic studies (X-rays, CT scans, MRIs, etc.), and other diagnostic procedures (e.g., EKG interpretations)
```

Identify and categorize any sensitive information in every claim according to these categories. If none of the above is identified, report "None."

**\*\*Format\*\*:**

**\*\*Input\*\*:**

- Claim 1: [Claim 1]
- Claim 2: [Claim 2]

**\*\*Output\*\*:**

- Claim 1: <Categories>
- Claim 2: <Categories>

**\*\*Input\*\*:**

{facts\_prepared}

**\*\*Output\*\*:**

### E.7.2 Categorize Sensitive Information - WildChat

Analyze the provided claims to identify segments containing sensitive information about individuals or groups.

#### ## Categories

There are 16 specific categories to consider:

- \* Age: Any mention of a person's age, e.g., "23-year-old"
- \* Gender: References to gender identity, e.g., "woman," "non-binary person"
- \* Sexual\_Orientation: Mentions of sexual orientation, e.g., "gay couple"
- \* Race\_Nationality: References to race, ethnicity, or nationality
- \* Spouse: Mentions of a person's wife, husband, or spouse
- \* Partner: References to a person's girlfriend, boyfriend, or partner
- \* Relationship\_Status: Mentions of marital status, being in a romantic relationship, or being single
- \* Family: References to family members or family structures
- \* Health: Includes a wide range of health-related information, from specific diseases or conditions to medications, medical tests, or treatments
- \* Mental\_Health: Includes a broad range of emotional states and mental health conditions, from feelings of sadness or anxiety to specific diagnoses
- \* Location: Captures specific geographical details about where a person lives or is located. Includes precise locations such as addresses, cities, countries, or distinctive landmarks
- \* Appearance: Physical descriptions of individuals, e.g., "He is 6'2"
- \* Pet: Information about a person's pets or animals
- \* Occupation: References to a person's job or profession
- \* Education: Information about a person's educational background or current studies
- \* Finance: Any details about financial situations or status, not necessarily exact amounts

Identify and categorize any sensitive information in every claim according to these categories. If none of the above is identified, report "None."

**\*\*Format\*\*:**

**\*\*Input\*\*:**

- Claim 1: [Claim 1]
- Claim 2: [Claim 2]

**\*\*Output\*\*:**

- Claim 1: <Categories>
- Claim 2: <Categories>

**\*\*Input\*\*:**

{facts\_prepared}

**\*\*Output\*\*:**