

# On the Vulnerability of Text Sanitization

Meng Tong<sup>1</sup>, Kejiang Chen<sup>1\*</sup>, Xiaojian Yuan<sup>1</sup>, Jiayang Liu<sup>2</sup>,  
Weiming Zhang<sup>1</sup>, Nenghai Yu<sup>1</sup>, and Jie Zhang<sup>3</sup>

<sup>1</sup>University of Science and Technology of China

<sup>2</sup>Nanyang Technological University

<sup>3</sup>CFAR and IHPC, A\*STAR, Singapore

{mtong,xjyuan}@mail.ustc.edu.cn, {chenkj,zhangwm,ynh}@ustc.edu.cn  
jiayang.liu@ntu.edu.sg, zhang\_jie@cfar.a-star.edu.sg

## Abstract

Text sanitization, which employs differential privacy to replace sensitive tokens with new ones, represents a significant technique for privacy protection. Typically, its performance in preserving privacy is evaluated by measuring the attack success rate (ASR) of reconstruction attacks, where attackers attempt to recover the original tokens from the sanitized ones. However, current reconstruction attacks on text sanitization are developed empirically, making it challenging to accurately assess the effectiveness of sanitization. In this paper, we aim to provide a more accurate evaluation of sanitization effectiveness. Inspired by the works of Palamidessi et al. (Alvim et al., 2015; Cherubin et al., 2019), we implement theoretically optimal reconstruction attacks targeting text sanitization. We derive their bounds on ASR as benchmarks for evaluating sanitization performance. For real-world applications, we propose two practical reconstruction attacks based on these theoretical findings. Our experimental results underscore the necessity of reassessing these overlooked risks. Notably, one of our attacks achieves a 46.4% improvement in ASR over the state-of-the-art baseline, with a privacy budget of  $\epsilon = 4.0$  on the SST-2 dataset. Our code is available at: <https://github.com/mengtong0110/On-the-Vulnerability-of-Text-Sanitization>.

## 1 Introduction

With the advancement of natural language processing (NLP) (Chang et al., 2023), concerns related to its privacy implications (Kim et al., 2024) have significantly intensified. For instance, Samsung employees previously leaked the company’s confidential meeting records by uploading documents to the cloud services of large language models (LLMs) (Sam, 2023). More recently, reports (Chowdhary, 2023; Wired, 2023) have dis-

\*Corresponding author.

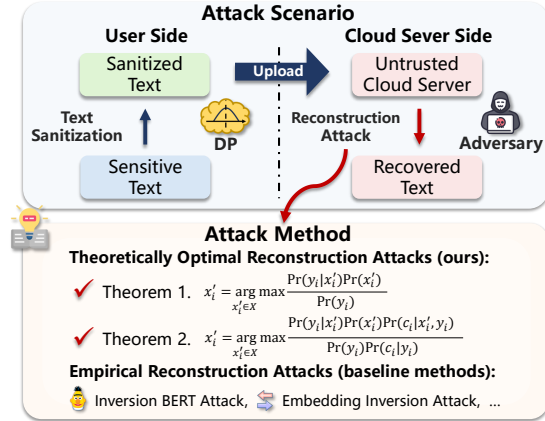


Figure 1: The illustration of reconstruction attacks.

closed a concerning flaw in LLM-based chatbots (OpenAI, 2023), which could expose user data and put private information at risk.

To address these privacy concerns, text sanitization (Chen et al., 2023; Meng et al., 2023) is now regarded as a crucial solution for privacy preservation in text and has been widely adopted by companies, such as Amazon (Amazon, 2024) and Microsoft (Microsoft, 2024b). Specifically, before users upload their documents to an untrusted cloud server for model training or inference, they can employ differential privacy (DP) (Dwork, 2006) to replace tokens corresponding to sensitive entities in the documents with newly sampled tokens. This method minimizes alterations in the uploaded data compared to other methods, such as synthetic data generation (Hong et al., 2024; Saiteja et al., 2023).

While DP sanitizes sensitive tokens, recent works (Zhou et al., 2023; Du et al., 2023) have demonstrated that attackers can still recover them from sanitized text. This finding underscores the significance of measuring the effectiveness of text sanitization. One of the principal methods for evaluating its effectiveness in privacy protection is through reconstruction attacks (Balle et al., 2022;

Yue et al., 2021), which aim to recover sensitive tokens from sanitized text, as illustrated in Figure 1.

**Existing Attacks.** In this context, several reconstruction attacks have been developed to evaluate the vulnerability of text sanitization: The *Embedding Inversion Attack* (Qu et al., 2021) reconstructs tokens by selecting the token with the closest embedding distance to the sanitized token (Song and Raghunathan, 2020). The *Inversion BERT Attack* (Kugler et al., 2021) trains an inversion model to map sanitized tokens back to their original forms. The *Mask Token Inference Attack* (Yue et al., 2021) replaces each token in the sanitized text with a “[MASK]” token to predict the raw one. However, these attacks are empirically developed, which makes it difficult to accurately measure the effectiveness of sanitization (as discussed in Section 2).

**Our Proposal.** In this paper, we aim to address the evaluation gap in privacy protection for DP-based text sanitization. Inspired by the works of Palamidessi et al. (Alvim et al., 2015; Cherubin et al., 2019), who estimate privacy leakage using Bayesian inference, we introduce both the *Context-free Optimal Reconstruction Attack* and the *Contextual Optimal Reconstruction Attack* for assessing text sanitization. We also derive theoretical bounds on the attack success rate (ASR), which serve as benchmarks for evaluating its performance in protecting privacy. To support real-world implementation, we propose two practical attacks based on these theoretical findings. Experimental results demonstrate that our reconstruction attacks achieve higher ASRs than baseline methods.

**Our Contributions.** We summarize our main contributions as follows:

1. We implement theoretically context-free and contextual optimal reconstruction attacks against text sanitization, deriving their bounds on ASR to serve as benchmarks for evaluating the effectiveness of the sanitization.
2. Based on these theoretical findings, we propose two practical reconstruction attacks: *Context-free Bayesian Attack* and *Contextual Bayesian Attack*. Notably, *Contextual Bayesian Attack* is the first to transform the reconstruction attack to a classification task.
3. We reassessed text sanitization methods across four datasets using seven reconstruction attacks. Our experimental results underscore the necessity of paying closer attention to these overlooked risks. Specifically, *Contextual Bayesian Attack* achieves

a 46.4% improvement in ASR over the state-of-the-art baseline at  $\epsilon = 4.0$  on the SST-2 dataset (Socher et al., 2013a).

## 2 Preliminaries and Related Work

**Text Sanitization.** Text sanitization (Chen et al., 2023; Yue et al., 2021) leverages differential privacy (DP) (Dwork, 2006) to replace sensitive tokens with newly sampled tokens. Specifically, let  $X$  represent the set of sensitive tokens and  $Y$  represent the set of sanitized tokens. An original sentence consists of sensitive tokens  $x_1, x_2, \dots, x_n$ , where  $x_i \in X$ . Its sanitized sentence is  $y_1, y_2, \dots, y_n$ , where the token  $y_i \in Y$  is independently sampled by DP with the probability  $\Pr(y_i|x_i)$ . Additionally, we define the sanitized context  $\mathbf{c}_i = y_1, y_2, \dots, y_{i-1}, y_{i+1}, \dots, y_n$  for the sanitized token  $y_i$  (or original token  $x_i$ ), where  $\mathbf{c}_i \in C$  and  $C$  denotes the set of sanitized contexts.

In addition to text sanitization, recent studies (Staab et al., 2024a,b) have demonstrated that LLMs can also protect private information.

**Definition of Sensitive Tokens.** In previous research (Meng et al., 2023; Papadopoulou et al., 2022; Plant et al., 2021), they have made varying choices on the definition of sensitive tokens. In our study, we set the sensitive tokens in accordance with the corresponding defense method of text sanitization (Chen et al., 2023; Yue et al., 2021).

**Reconstruction Attacks.** The *Embedding Inversion Attack* (Qu et al., 2021) identifies the token closest to a sanitized token by comparing their embedding distances (Song and Raghunathan, 2020) as the reconstructed token. The *Inversion BERT Attack* (Kugler et al., 2021) uses a trained model to convert sanitized representations back to their original tokens on a one-to-one basis. However, neither method has performed a comprehensive evaluation of sanitization effectiveness, which should consider the context of the sanitized tokens. The *Mask Token Inference Attack* (Yue et al., 2021) replaces each token in sanitized text with a “[MASK]” token, which is then predicted to reconstruct the raw token. This attack overlooks the impact of probability distributions of raw tokens and the sampling probabilities in DP on the attack result, leading to an inaccurate evaluation of sanitization performance.

To address these issues, we introduce optimal reconstruction attacks and propose corresponding practical approaches for evaluating the effectiveness of DP-based text sanitization.

### 3 Theoretically Optimal Reconstruction Attacks

In this section, we first present the threat model and then develop *Context-free Optimal Reconstruction Attack*, deriving its ASR bound to enable quick evaluation of text sanitization. For a comprehensive evaluation, we propose *Contextual Optimal Reconstruction Attack* with its ASR bound.

#### 3.1 Threat Model for Theoretical Attacks

**Adversary’s Goal.** The goal of an adversary is to obtain information on sensitive tokens by reconstructing each original token from its sanitized one.

**Adversary’s Knowledge.** We assume an adversary who knows the differentially private algorithm with its sampling probabilities  $\Pr(Y|X)$ , which is utilized in text sanitization. Additionally, we propose an assumption that the adversary possesses the following theoretical probabilities:  $\Pr(X)$  (the probability of a token appearing in an original sentence) and  $\Pr(C|X, Y)$  (the probability that a sanitized context corresponds to an original token and a sanitized token in the sanitization process).

#### 3.2 Warm-up: Context-free Optimal Reconstruction Attack

Building on the Bayesian estimation of DP (Alvim et al., 2015; Cherubin et al., 2019), we introduce the context-free optimal strategy for reconstructing the original tokens from sanitized tokens. We further derive its tight ASR bound, which achieves a quick evaluation of sanitization effectiveness.

**Theorem 1 (Context-free Optimal Reconstruction Attack).** *Given a sanitized token  $y \in Y$ , the optimal strategy to reconstruct the original token of  $y$  is to select the token  $x' \in X$  according to the following rule:*

$$x' = \arg \max_{x' \in X} \Pr(x'|y) \quad (1)$$

$$= \arg \max_{x' \in X} \frac{\Pr(y|x') \Pr(x')}{\Pr(y)}. \quad (2)$$

The proof of Theorem 1 is in Appendix A. It is shown that an adversary should select the most likely original token  $x'$  corresponding to the sanitized token  $y$  as the reconstructed token. According to Bayes’ formula (Joyce, 2003), Equation (1) can be decomposed into several components.  $\Pr(y)$  is constant for all  $x' \in X$  and does not affect the outcome of the attack. For  $\Pr(y|x')$  and  $\Pr(x')$ ,

it is observed that: 1) A token  $x'$  with a higher sampling probability  $\Pr(y|x')$  is more likely to be the original token of  $y$ ; 2) A token  $x'$  with a higher  $\Pr(x')$  appears more frequently and is more likely to be the raw token of  $y$ .

Based on the context-free optimal reconstruction attack, we derive its tight bound on the ASR as a quick evaluation for the effectiveness of DP-based text sanitization:

**Definition 1 (Context-free ASR Bound).** *Let  $x_1^j, x_2^j, \dots, x_{n_j}^j$  and  $y_1^j, y_2^j, \dots, y_{n_j}^j$  represent the  $j$ -th original sentence and its sanitized sentence by DP. Given any  $m$  sanitized sentences, the Context-free Attack Success Rate Bound (Context-free ASR Bound) is defined as follows:*

$$\text{Context-free ASR Bound} := \frac{\sum_{j=1}^m \sum_{i=1}^{n_j} \delta_{x_i^j, x_i^{j'}}}{\sum_{i=1}^m n_j}, \quad (3)$$

$$x_i^{j'} = \arg \max_{x_i^{j'} \in X} \Pr(x_i^{j'}|y_i^j), \delta_{x_i^j, x_i^{j'}} = \begin{cases} 1, & x_i^j = x_i^{j'} \\ 0, & x_i^j \neq x_i^{j'} \end{cases}. \quad (4)$$

The *Context-free ASR Bound* serves as a tight ASR bound for the *Context-free Optimal Reconstruction Attack*. This is because the attack of *Context-free ASR Bound* strictly follows Theorem 1 to reconstruct the sanitized tokens. This bound primarily utilizes simple frequency statistics, making it a convenient tool for swiftly evaluating DP-based text sanitization methods. To represent  $\Pr(x_i^{j'})$ , which is informed to the adversary in the threat model, it computes the theoretical probability of the original token  $x_i^{j'}$  appearing in the original sentences. The term  $\Pr(y_i^j|x_i^{j'})$  denotes the probability that DP samples sanitized token  $y_i^j$  with original token  $x_i^{j'}$ . Although *Context-free ASR Bound* swiftly assesses the sanitization effectiveness, it does not consider sanitized context, which may lead to a less comprehensive evaluation. Therefore, we develop *Contextual Optimal Reconstruction Attack*.

#### 3.3 Advanced: Contextual Optimal Reconstruction Attack

**Theorem 2 (Contextual Optimal Reconstruction Attack).** *Given a sanitized token  $y$  and its sanitized context  $c$ , the optimal strategy to reconstruct the original token of  $y$  is to select the token  $x' \in X$*

according to the following rule:

$$x' = \arg \max_{x' \in X} \Pr(x'|y, c) \quad (5)$$

$$= \arg \max_{x' \in X} \frac{\Pr(y|x') \Pr(x') \Pr(c|x', y)}{\Pr(y) \Pr(c|y)}. \quad (6)$$

The proof of Theorem 2 is provided in Appendix A. It demonstrates that an adversary should select the most probable original token  $x'$ , corresponding to the sanitized token  $y$  and its sanitized context  $c$  (previously defined in Section 2), as the reconstructed token. Theorem 2 reduces the amount of uncertainty in reconstructing the original tokens compared to Theorem 1. This holds because  $H(X|Y, C) \leq H(X|Y)$ , where  $H(X|Y)$  is the entropy of  $X$  conditioned on  $Y$ , and  $H(X|Y, C)$  is the entropy of  $X$  conditioned on both  $Y$  and  $C$ . Additionally, Theorem 2 includes extra calculations for  $\Pr(c|x', y)$  and  $\Pr(c|y)$  relative to Theorem 1.  $\Pr(c|y)$  remains constant for all  $x' \in X$  and does not influence the outcome of the attack.  $\Pr(c|x', y)$  represents the probability that  $c$  is the sanitized context of both  $x'$  and  $y$ .

**Challenge of Computation.** However, since  $c$  is a sequence that consists of sanitized tokens, accurately calculating the theoretical value of  $\Pr(c|x', y)$  poses a challenge. Due to this intractability, we use its approximation to derive the ASR bound of attack strategy in Theorem 2.

**Definition 2 (Contextual  $K$ -ASR Bound).** Let  $x_1^j, x_2^j, \dots, x_{n_j}^j$  and  $y_1^j, y_2^j, \dots, y_{n_j}^j$  represent the  $j$ -th original sentence and its sanitized sentence by DP. Let the sanitized context  $c_i^j$  for the  $i$ -th sanitized token  $y_i^j$  (whose original token is  $x_i^j$ ) be  $c_i^j = y_1^j, y_2^j, \dots, y_{i-1}^j, y_{i+1}^j, \dots, y_{n_j}^j$ . Given any  $m$  sanitized sentences, the Contextual  $K$  Attack Success Rate Bound (Contextual  $K$ -ASR Bound) is defined as follows:

$$\text{Contextual } K\text{-ASR Bound} := \frac{\sum_{j=1}^m \sum_{i=1}^{n_j} \delta_{x_i^j, x_i^{j'}}}{\sum_{j=1}^m n_j}, \quad (7)$$

$$x_i^{j'} = \arg \max_{x_i^{j'} \in X_i^{j'}} \Pr(x_i^{j'}|y_i^j, c_i^j), \quad (8)$$

$$X_i^{j'} = \arg \text{top } K \Pr(x_i^{j'}|y_i^j), \delta_{x_i^j, x_i^{j'}} = \begin{cases} 1, & x_i^j = x_i^{j'} \\ 0, & x_i^j \neq x_i^{j'} \end{cases}. \quad (9)$$

The Contextual  $K$ -ASR Bound provides an approximation of the ASR bound derived from Theorem 2. It calculates only the top  $K$  values of

$\Pr(x_i^{j'}|y_i^j, c_i^j)$ , sorted by  $\Pr(x_i^{j'}|y_i^j)$ . Because traversing the entire  $x_i^{j'} \in X$  for the attack outcome requires a significant amount of time, as detailed in Appendix B. We find that it is not necessary to do so. The experiments in Section 5.4 indicate that the top  $K$  tokens sorted by  $\Pr(x_i^{j'}|y_i^j)$  include the majority of original tokens. Increasing  $K$  does not necessarily improve ASR. This method balances the trade-off between time cost and ASR.

For probability computations in the Contextual  $K$ -ASR Bound, it easily calculates  $\Pr(x_i^{j'})$  based on the theoretical probability of the original token  $x_i^{j'}$  appearing in the original sentences. However, given that  $c_i^j$  is a sequence of sanitized tokens, it is challenging to calculate probability  $\Pr(c_i^j|x_i^{j'}, y_i^j)$ . Although the chain rule can break it down into multiple components (Bain and Engelhardt, 1992), rebuilding the conditional probabilities for each  $c_i^j$  is time-consuming.

To approximate the probability  $\Pr(c_i^j|x_i^{j'}, y_i^j)$ , we convert it into a binary classification task. We represent it by the prediction probability of a specific label from a BERT model (Devlin et al., 2019) that has been fine-tuned on both the sanitized and original sentences. We detail the implementation of this binary classification in the following section 4.

Based on the preceding discussion, we have outlined a panoramic view of the theoretically optimal reconstruction attacks and their corresponding ASR bounds. It is important to note that these attacks are developed based on theoretical probabilities. However, practical reconstruction attacks often fall short of this level of precision, typically approximating probability distributions using a shadow dataset. To address this, we introduce the Context-free Bayesian Attack and the Contextual Bayesian Attack as practical implementations of theoretically optimal reconstruction attacks.

## 4 Practical Implementations of Theorems

### 4.1 Threat Model for Practical Attacks

**Adversary's Goal.** The goal of an adversary is to obtain sensitive information by reconstructing each original token from its sanitized token and corresponding sanitized context.

**Adversary's Knowledge.** We assume an adversary that accesses the information: 1) *Differentially Private Algorithms*. We assume that an adversary is informed about the differentially private algorithms with its sampling probabilities  $\Pr(Y|X)$  for text



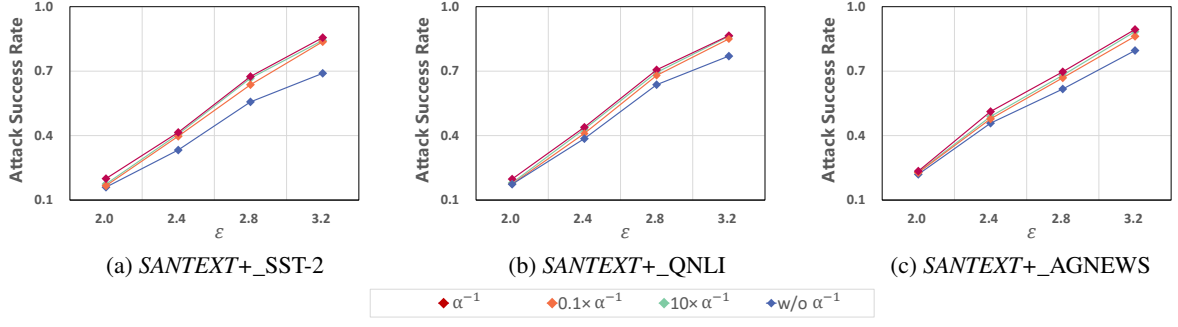


Figure 2: ASR of *Context-free Bayesian Attack* using various constants instead of  $\alpha^{-1}$  on SST-2 dataset.

sanitization. Differentially private algorithms are typically treated as public access (Kairouz et al., 2014; Cormode et al., 2018; Wang et al., 2019c; Murakami and Kawamoto, 2019). 2) *A Shadow Dataset*. Additionally, we assume that an adversary has access to a shadow dataset, consistent with related privacy attacks (Carlini et al., 2022; Ye et al., 2022). The shadow dataset is related to but disjointed from the original sentences corresponding to the sanitized tokens under attack.

## 4.2 Context-free Bayesian Attack (based on Theorem 1)

**Attack Strategy in Context-free Bayesian Attack.** Given a sanitized token  $y \in Y$ , the strategy for Context-free Bayesian Attack to reconstruct the original token of  $y$  is to select the token  $x' \in X$  according to the following rule:

$$x' = \arg \max_{x' \in X} \frac{\Pr(y|x') (\Pr(x') + \alpha^{-1})}{\Pr(y)}, \quad (10)$$

where  $\Pr(x')$  is the statistical probability in the shadow dataset, and  $\alpha$  is the number of tokens in the shadow dataset.

Unfortunately, the practical implementation of Theorem 1 encounters difficulty when the approximate probabilities  $\Pr(x')$  of all potential original tokens corresponding to  $y$  are zero in the shadow dataset. Under these conditions, it is challenging to select an  $x'$  that uniquely maximizes  $\Pr(x'|y)$ , as these probabilities  $\Pr(x'|y)$  are all zero. To address this, a small constant,  $\alpha^{-1}$ , is introduced in the attack. This modification enables *Context-free Bayesian Attack* to select  $x' = \arg \max \Pr(y|x')$  as the reconstructed token in such cases. Moreover,  $\alpha^{-1}$  has little effect on the outcome of the reconstruction attack when the probability  $\Pr(x')$  is not zero.

Figure 2 presents the ablation studies on parameter  $\alpha^{-1}$  with the *SANTEXT+* (Yue et al., 2021) defense, which demonstrated the effectiveness of  $\alpha^{-1}$  in practical reconstruction attacks. More experiments discussed in Appendix C further explain why we use  $\alpha^{-1}$  in *Context-free Bayesian Attack*.

## 4.3 Contextual Bayesian Attack (based on Theorem 2)

**Attack Strategy in Contextual Bayesian Attack.**

Given a sanitized token  $y$  and its sanitized context  $c$ , the strategy for Contextual Bayesian Attack to reconstruct the original token of  $y$  is to select the token  $x' \in X$  according to the following rule:

$$x' = \arg \max_{x' \in X'} \frac{\Pr(y|x') \cdot (\Pr(x') + \alpha^{-1})}{\Pr(y)} \times \frac{\Pr(c|x', y)}{\Pr(c|y)}, \quad (11)$$

$$X' = \arg \top K_{x' \in X} \frac{\Pr(y|x') \cdot (\Pr(x') + \alpha^{-1})}{\Pr(y)}, \quad (12)$$

where  $\Pr(x')$  is the statistical probability in shadow dataset;  $\alpha$  is the count of tokens in shadow dataset;  $\Pr(c|x', y)$  is obtained from a fine-tuned model on shadow dataset.

**Transforming Reconstruction to Classification.**

To represent the probability  $\Pr(c|x', y)$ , we transform it with constant tokens  $x'$  and  $y$ :  $\Pr(c|x', y) = \Pr(f(c, x'), f(c, y) | x', y)$ , where the function  $f$  is defined as follows:

Given any  $c = y_1, y_2, \dots, y_{i-1}, y_{i+1}, \dots, y_n$ , let  $f(c, x') = y_1, y_2, \dots, y_{i-1}, x', y_{i+1}, \dots, y_n$  and  $f(c, y) = y_1, y_2, \dots, y_{i-1}, y, y_{i+1}, \dots, y_n$ .

To calculate  $\Pr(f(c, x'), f(c, y) | x', y)$ , we first fine-tune a BERT model (Devlin et al., 2019) to serve as a detector, denoted by  $h$ . This detector

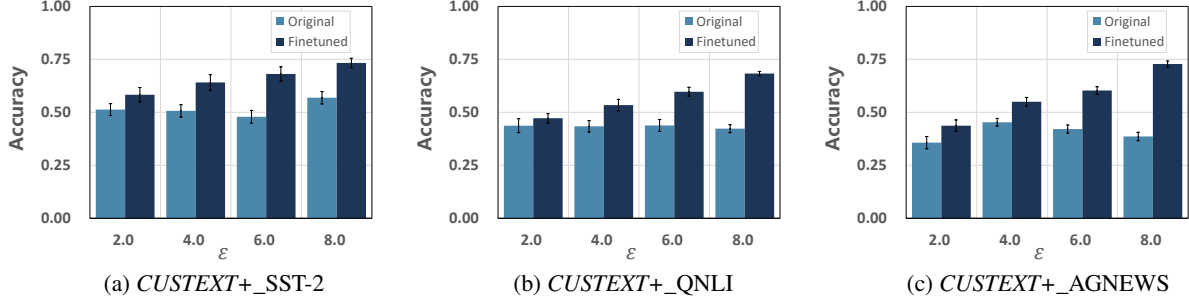


Figure 3: Classification accuracy of detector  $h$  in *Contextual Bayesian Attack* with CUSTEXT+.

$h$  is a binary classifier that infers whether  $f(c, x')$  and  $f(c, y)$  correspond to  $x'$  and  $y$  in sanitization process (i.e.,  $c$  is the sanitized context for  $x'$  and  $y$ ). Specifically, the input to detector  $h$  consists of two token sequences,  $f(c, x')$  and  $f(c, y)$ . The output of detector  $h$  is either label 1 or label 0, where a value of 1 indicates that  $f(c, x')$  and  $f(c, y)$  correspond to  $x'$  and  $y$ , while a value of 0 indicates that they do not. With these inputs and outputs,  $h$  is defined as  $h(f(c, y), f(c, x')) \rightarrow \{0, 1\}$ . The training samples of  $h$  are constructed as follows:

**Step 1: Confirm the Attack Target.** We first confirm the attack target  $\Gamma$  in order to construct corresponding training samples. The attack target  $\Gamma$  comprises a set of sanitized tokens that need to be reconstructed to their original forms.

**Step 2: Construct the Training Samples.** We sanitize sentences in the shadow dataset using the differentially private algorithm. For each token  $y \in \Gamma$  in the sanitized sentences, we create a training sample consisting of two token sequences and a label for  $y$ . These two token sequences are the input of detector  $h$ : The first sequence includes the sanitized sentence containing  $y$  and  $c$ . The second sequence replaces the token  $y \in \Gamma$  in the first sequence with its reconstructed result  $x'$  from the *Context-free Bayesian Attack*. The label is the output of detector  $h$ : If  $x'$  matches the original token of  $y$ , indicating that  $f(c, x')$  and  $f(c, y)$  correspond to  $x'$  and  $y$  in sanitization process, we label the sample 1. Otherwise, we label it 0. To maintain label balance, for each sample labeled as 1, we generate a corresponding sample labeled as 0 using the same sanitized sentence but with the second most likely reconstructed token from the *Context-free Bayesian Attack*. Similarly, for each sample labeled 0, we generate a sample labeled 1.

With the constructed training samples, we fine-tune the BERT model over three epochs with

*Adam Optimizer* (Kingma and Ba, 2015), obtaining the detector  $h$ . We represent  $\Pr(c|x', y) = \Pr(f(c, x'), f(c, y) | x', y)$  using prediction probability of  $h$  for label 1.

Figure 3 shows the classification accuracy of  $h$ , as detailed in Appendix D. Specifically, “Original” in Figure 3 refers to the performance of directly using BERT without any fine-tuning, while “Fine-tuned” refers to the performance after fine-tuning BERT on the detection task. Experimental results demonstrate the effectiveness of  $h$  in determining whether  $c$  is the sanitized context of  $x'$  and  $y$ .

## 5 Experiments

### 5.1 Experimental Setup

**Datasets.** To validate the effectiveness of our attacks, we conduct experiments on four datasets, SST-2 (Socher et al., 2013a), AGNEWS (Zhang et al., 2015a), QNLI (Wang et al., 2019a), and Yelp (Zhang et al., 2015a). These datasets cover a wide range of sensitive entities and personally identifiable information (PII) (detailed in Appendix E).

**Defense Methods.** We implement reconstruction attacks against two backbone methods of text sanitization using their default settings for sensitive tokens and parameters: CUSTEXT+ (Chen et al., 2023) and SANTEXT+ (Yue et al., 2021), both of which are widely used in DP research.

**Baseline Attacks.** Our work presents the first systematic study of reconstruction attacks against text sanitization. Since there is no other research specifically focusing on reconstruction attacks against text sanitization, we compare our methods with three adaptive attacks from prior defenses (Meng et al., 2023; Zhou et al., 2023): *Embedding Inversion Attack* (Qu et al., 2021), *Inversion BERT Attack* (Kugler et al., 2021), and *Mask Token Inference Attack* (Yue et al., 2021).

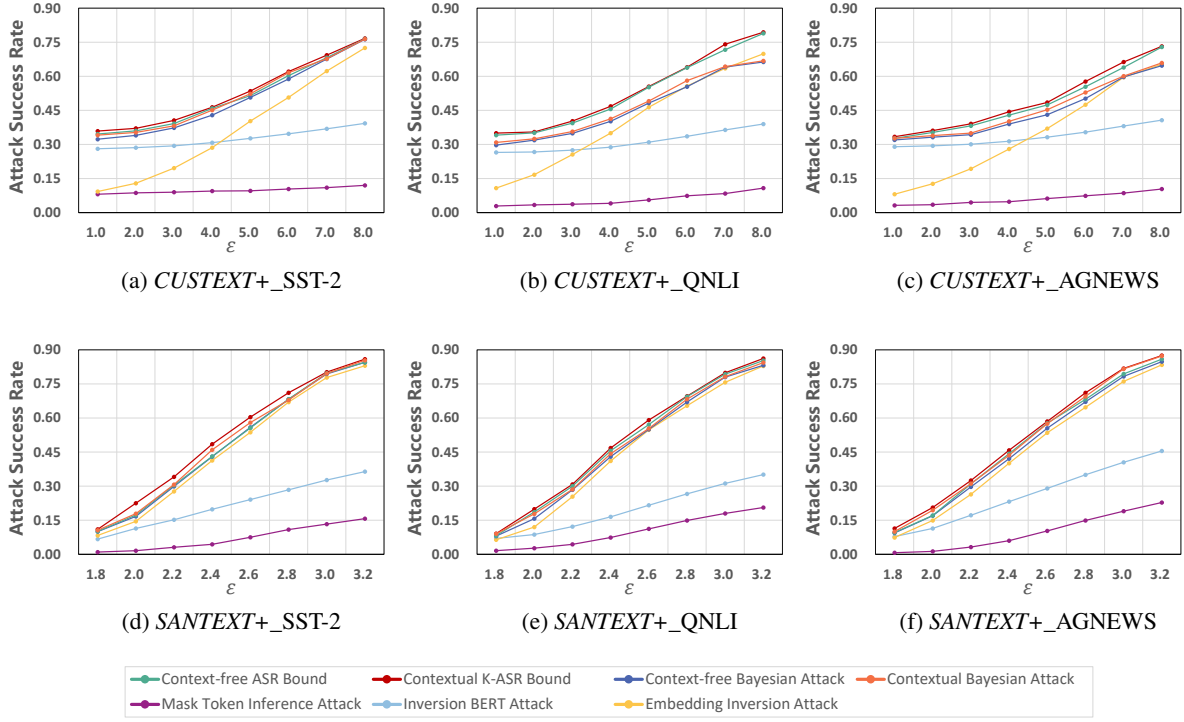


Figure 4: ASR of reconstructing sensitive tokens against CUSTEXT+ and SANTEXT+ with  $K = 10$ .

**Implementation.** We run experiments on a RTX A6000 GPU. We randomly select 20,000 sentences from the entire dataset to form a private dataset, whose sensitive tokens are perturbed to sanitized tokens. In CUSTEXT+, we calculate the ASR of the successful reconstruction for the sensitive tokens from 3,000 sanitized tokens. In SANTEXT+, this calculation of ASR is based on 1,000 sanitized tokens. For a shadow dataset, we randomly select 1,000 sentences from the entire dataset, which are disjoint from those to reconstruct.

More implementation details of our experiments can be found in Appendix F.

## 5.2 Performance of Reconstruction Attacks

**Baseline Reconstruction Attacks.** In this paper, we identify three factors that facilitate reconstructing original tokens: the sampling probabilities of DP, the probability distribution of input tokens, and the sanitized context. However, the baseline attacks generally fall short in a design that considers all three factors. The limitations of them are demonstrated by the experimental results. For instance, at  $\epsilon = 3.2$  on the QNLI, as shown in Figure 4e, both Inversion BERT Attack and Mask Token Inference Attack successfully reconstructed the sanitized to-

kens less than half of those in others. Moreover, at  $\epsilon = 1.0$  on the AGNEWS, as shown in Figure 4c, the ASR of Embedding Inversion Attack is less than 0.09, which is approximately one-third of that in Inversion BERT Attack.

**Attack Success Rate Bounds.** Experimental results in Figure 4 indicate that both Context-free ASR Bound and Contextual K-ASR Bound more accurately evaluate the effectiveness of DP-based text sanitization compared to baselines. For instance, with a privacy parameter  $\epsilon = 1.0$  on QNLI, as shown in Figure 4b, the ASR of the Context-free ASR Bound is 28.7% higher than that of the Inversion BERT Attack, while the ASR of the Contextual K-ASR Bound is 32.1% higher than that of the Inversion BERT Attack. Furthermore, our experiments demonstrate that the Contextual K-ASR Bound is consistently higher than the Context-free ASR Bound. Specifically, as shown in Figure 4d, the Contextual K-ASR Bound exceeds the Context-free ASR Bound by 30.1% with  $\epsilon = 2.0$  on SST-2.

**Practical Attacks from Theorems.** According to Figure 4, the ASR of both the Context-free Bayesian Attack and the Contextual Bayesian Attack exceeded that of the baseline attacks in most cases. Notably, the Contextual Bayesian Attack

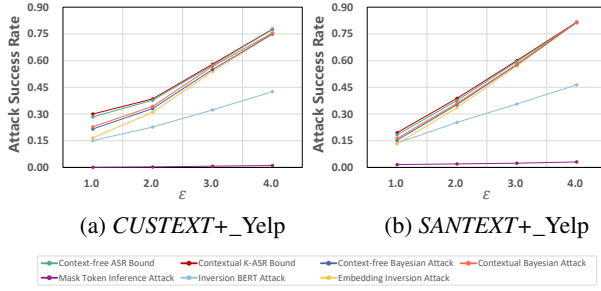


Figure 5: ASR of reconstructing PII tokens.

Table 1: ASR of reconstruction attacks with various  $K$ .

Attack Strategy	K=1	K=5	K=10	K=15	K=20
Contextual Bound	0.499	0.520	0.525	0.525	0.525
Contextual Attack	0.470	0.497	0.503	0.503	0.503

achieves a 46.4% improvement in ASR over the *Inversion BERT Attack* at a privacy budget of  $\epsilon = 4.0$  on the SST-2 according to Figure 4a. With the same experimental setting, the *Context-free Bayesian Attack* achieves a 39.2% improvement in ASR over the *Inversion BERT Attack*. Furthermore, the ASRs of our proposed practical attacks are approaching their corresponding theoretical ASR bounds. For example, with  $\epsilon = 1.8$  on SST-2 as shown in Figure 4d, the ASR of the *Context-free Bayesian Attack* was only 0.001 less than that of the *Context-free ASR Bound*. With the same setting, the ASR of the *Contextual Bayesian Attack* was 0.004 less than that of the *Contextual K-ASR Bound*. More discussion on the performance of reconstruction attacks with large  $\epsilon$  can be found in Appendix H.

Furthermore, our reconstruction attacks remain comparable to the baseline methods even with a misaligned shadow dataset. The detailed experiments can be found in Appendix I.

### 5.3 Reconstructing PII Tokens

Text sanitization is widely adopted in de-identifying personal information (Amazon, 2024; Microsoft, 2024b). To bring the evaluations closer to real-life scenarios, we conducted reconstruction attacks on the Yelp dataset to recover the PII of personal names (Microsoft, 2024a) from sanitized sentences. Experimental results, as illustrated in Figure 5, indicate the improvement of our attacks over the baselines. For instance, as shown in Figure 5a, the *Context-free ASR Bound* increased the ASR by 41.5% with a privacy parameter  $\epsilon = 1.0$  over *Em-*

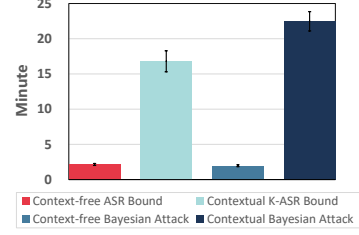


Figure 6: Time costs of reconstructing 1,000 tokens.

Table 2: ASR of reconstructing sensitive tokens with different shadow(S) / private(P) dataset size ratios.

Attack Strategy	S/P=1 %	S/P=3 %	S/P=5 %	S/P=8 %
Context-free Attack	0.492	0.494	0.496	0.506
Contextual Attack	0.473	0.487	0.505	0.508

*bedding Inversion Attack*; at the same experimental setting, *Contextual K-ASR Bound* increased ASR by 96.1% over *Inversion BERT Attack*.

### 5.4 Time Cost and Parameter Influence

**Time Costs.** We estimated the time required for reconstructing 1,000 sanitized tokens. We conducted this experiment 10 times to calculate standard deviations (1-sigma) using the closed form formula (Altman and Bland, 2005). As depicted in Figure 6, computations for the *Context-free ASR Bound* and the *Context-free Bayesian Attack* were completed in less than 5 minutes. Both of them enable quick evaluation of the effectiveness of text sanitization. For a comprehensive evaluation, *Contextual Bayesian Attack* and *Contextual K-ASR Bound* completed reconstructions in less than 25 minutes. We further compare the time costs of various reconstruction attacks in Appendix G.

**Parameter Influences.** Additionally, we test the influence of parameters on our attacks using SST-2 and CUSTEXT+ at  $\epsilon = 4.0$ . Table 1 shows the influence of  $K$  in both *Contextual Bayesian Attack* (Contextual Attack) and *Contextual K-ASR Bound* (Contextual Bound). When  $K$  is smaller than 10, the ASR of both grows with  $K$ . When  $K$  exceeds 10, the growth of ASR tends to flatten. Table 2 shows the ASRs for different size ratios between the shadow dataset and the private dataset. With a smaller ratio, the ASR of *Contextual Bayesian Attack* (Contextual Attack) and *Context-free Bayesian Attack* (Context-free Attack) decreases.



## 6 Conclusion

In this work, we implement both context-free and contextual optimal reconstruction attacks against text sanitization. We derive their bounds on ASR to guide the evaluation of sanitization effectiveness. We propose two practical attacks for real-world implementations. Evaluation results validate that our attacks surpass baseline methods.

This work represents an initial attempt to provide a more precise evaluation of sanitization effectiveness compared to current empirical attacks. Through this endeavor, we hope our work contributes to the research of privacy-preserving NLP.

## 7 Limitations

The results presented in our paper have two main limitations. First, we have not found a method to calculate the tight ASR bound of the *Contextual Optimal Reconstruction Attack* due to the challenge of accurately computing the theoretical value of  $\Pr(c|x', y)$ . Therefore, the calculation of this tight ASR bound remains an open question. Second, the *Contextual Bayesian Attack* entails model fine-tuning and inference. There is still room for a more time-efficient reconstruction attack.

Furthermore, recent studies on text anonymization (Staab et al., 2024a,b), which leverage advanced LLMs to sanitize sensitive tokens in text, present an exciting avenue for future research. However, the complexity of LLM-based methods introduces unique challenges that require careful consideration. Consequently, the privacy guarantees of these methods are not explored in this paper. We believe that investigating these techniques could provide valuable insights and represent an important direction for future work.

## Acknowledgments

This research is supported by the National Natural Science Foundation of China under Grant 62472398, Grant U2336206, and Grant U2436601, as well as by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative (No. DTCRGC-04). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority.

## References

2023. Samsung employees leak proprietary information via chatgpt. <https://www.cshub.com/data/news/iotw-samsung-employees-allegedly-leak-proprietary-information-via-chatgpt>.
- Douglas G Altman and J Martin Bland. 2005. Standard deviations and standard errors. *Bmj*, 331(7521):903.
- Mário S Alvim, Miguel E Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. 2015. On the information leakage of differentially-private mechanisms. *Journal of Computer Security*, 23(4):427–469.
- Amazon. 2024. Discover and protect your sensitive data at scale. <https://aws.amazon.com/maciel/>.
- Lee J Bain and Max Engelhardt. 1992. *Introduction to probability and mathematical statistics*, volume 4. Duxbury Press Belmont, CA.
- Borja Balle, Giovanni Cherubin, and Jamie Hayes. 2022. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1138–1156. IEEE.
- Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.
- Sai Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. 2023. A customized text sanitization mechanism with differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5747–5758.
- Giovanni Cherubin, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2019. F-bleau: fast black-box leakage estimation. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 835–852. IEEE.
- Krishi Chowdhary. 2023. Custom openai chatbots leak secrets. <https://techreport.com/news/custom-openai-chatbots-leak-secrets-pose-privacy-threats/>.
- Graham Cormode, Somesh Jha, Tejas Kulkarni, Ninghui Li, Divesh Srivastava, and Tianhao Wang. 2018. Privacy at scale: Local differential privacy in practice. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1655–1658.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

- Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Minxin Du, Xiang Yue, Sherman SM Chow, Tianhao Wang, Chenyu Huang, and Huan Sun. 2023. Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 2665–2679.
- Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.
- Junyuan Hong, Jiachen T. Wang, Chenhui Zhang, Zhangheng LI, Bo Li, and Zhangyang Wang. 2024. [DP-OPT: Make large language model your privacy-preserving prompt engineer](#). In *The Twelfth International Conference on Learning Representations*.
- James Joyce. 2003. Bayes’ theorem.
- Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2014. Extremal mechanisms for local differential privacy. *Advances in neural information processing systems*, 27.
- Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2024. Propile: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems*, 36.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Kai Kugler, Simon Münker, Johannes Höhmann, and Achim Rettinger. 2021. Invert: Reconstructing text from contextualized word embeddings by inverting the bert pipeline. *arXiv preprint arXiv:2109.10104*.
- Tong Meng, Chen Kejiang, Zhang Jie, Qi Yuang, Zhang Weiming, Yu Nenghai, Zhang Tianwei, and Zhang Zhikun. 2023. Inferdpt: Privacy-preserving inference for black-box large language models. *arXiv preprint arXiv:2310.12214*.
- Microsoft. 2024a. Data protection and de-identification sdk. <https://microsoft.github.io/presidio/>.
- Microsoft. 2024b. Modernize privacy program with a platform your. <https://www.microsoft.com/en-us/security/business/microsoft-privva>.
- Takao Murakami and Yusuke Kawamoto. 2019. {Utility-optimized} local differential privacy mechanisms for distribution estimation. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1877–1894.
- OpenAI. 2023. We’re launching the gpt store to help you find useful and popular custom versions of chatgpt. <https://openai.com/blog/introducing-the-gpt-store>.
- Anthi Papadopoulou, Yunhao Yu, Pierre Lison, and Lilja Øvrelid. 2022. Neural text sanitization with explicit measures of privacy risk. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 217–229.
- Richard Plant, Dimitra Gkatzia, and Valerio Giuffrida. 2021. [CAPE: context-aware private embeddings for private language learning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7970–7978. Association for Computational Linguistics.
- Chen Qu, Weize Kong, Liu Yang, Mingyang Zhang, Michael Bendersky, and Marc Najork. 2021. Natural language understanding with privacy-preserving bert. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1488–1497.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [Squad: 100, 000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.
- Utpala Saiteja, Hooker Sara, and Chen Pin-Yu. 2023. [Locally differentially private document generation using zero shot prompting](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8442–8457, Singapore. Association for Computational Linguistics.
- Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013a. [Recursive deep models for semantic compositionality over a sentiment treebank](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013b. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

- Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 377–390.
- Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024a. Large language models are advanced anonymizers. *arXiv preprint arXiv:2402.13846*.
- Robin Staab, Mark Vero, Mislav Balunovic, and Martin T. Vechev. 2024b. [Beyond memorization: Violating privacy via inference with large language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. [SKEP: sentiment knowledge enhanced pre-training for sentiment analysis](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4067–4076. Association for Computational Linguistics.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *International Conference on Learning Representations*.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Ning Wang, Xiaokui Xiao, Yin Yang, Jun Zhao, Siu Cheung Hui, Hyejin Shin, Junbum Shin, and Ge Yu. 2019c. Collecting and analyzing multidimensional data with local differential privacy. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 638–649. IEEE.
- Wired. 2023. Openai’s custom chatbots are leaking their secrets. <https://www.wired.com/story/openai-custom-chatbots-gpts-prompt-injection-attacks/>.
- Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. 2022. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 3093–3106.
- Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. 2021. [Differential privacy for text analytics via natural text sanitization](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3853–3866, Online. Association for Computational Linguistics.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015a. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015b. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.
- Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.
- Xin Zhou, Yi Lu, Ruotian Ma, Tao Gui, Yuran Wang, Yong Ding, Yibo Zhang, Qi Zhang, and Xuan-Jing Huang. 2023. Textobfuscator: Making pre-trained language model a privacy protector via obfuscating word representations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5459–5473.

## A Proof of Theorems

**Theorem 1 (Context-free Optimal Reconstruction Attack).** *Given a sanitized token  $y \in Y$ , the optimal strategy to reconstruct the original token of  $y$  is to select the token  $x' \in X$  according to the following rule:*

$$x' = \arg \max_{x' \in X} \Pr(x'|y) \quad (13)$$

$$= \arg \max_{x' \in X} \frac{\Pr(y|x') \Pr(x')}{\Pr(y)}. \quad (14)$$

*Proof of Theorem 1.* We aim to prove that the attack in Theorem 1 maximizes the expected value of the reconstructed token matching the original token for  $y$ . Let  $X$  denote the set of probable original tokens for  $y$ , defined as  $X = \{x_1, x_2, \dots, x_i, \dots, x_N\}$ , where  $x_i \in X$  is a probable original token of  $y$  with probability  $\Pr(x_i|y)$ . Let the probability that an attacker infers  $x_i \in X$  as the reconstructed token of  $y$  be  $p_i$ . It holds that  $\sum_{i=1}^N p_i = 1$ . The expected value of the reconstructed token matching the original token of  $y$  can be expressed as:

$$\Pr(x_1|y) p_1 + \Pr(x_2|y) p_2 + \dots + \Pr(x_N|y) p_N. \quad (15)$$

With  $\sum_{j=1}^N p_{ij} = 1$ , it can be deduced that:

$$\Pr(x_1|y) p_1 + \Pr(x_2|y) p_2 + \dots + \Pr(x_N|y) p_N \leq \max_{x' \in X} \Pr(x'|y). \quad (16)$$

According to Equation (16), the expected value of correctly reconstructing the original token is smaller than  $\max_{x' \in X} \Pr(x'|y)$ , which is the expected value under attack strategy in Theorem 1.  $\square$

**Theorem 2 (Contextual Optimal Reconstruction Attack).** *Given a sanitized token  $y$  and its sanitized context  $\mathbf{c}$ , the optimal strategy to reconstruct the original token of  $y$  is to select the token  $x' \in X$  according to the following rule:*

$$\begin{aligned} x' &= \arg \max_{x' \in X} \Pr(x'|y, \mathbf{c}) \\ &= \arg \max_{x' \in X} \frac{\Pr(y|x') \Pr(x') \Pr(\mathbf{c}|x', y)}{\Pr(y) \Pr(\mathbf{c}|y)}. \end{aligned} \quad (17)$$

$$(18)$$

*Proof of Theorem 2.* We aim to prove that the attack in Theorem 2 maximizes the expected value of the reconstructed token matching the original token for  $y$  and  $\mathbf{c}$ . Let  $X$  denote the set of probable original tokens for  $y$  and  $\mathbf{c}$ , defined as  $X = \{x_1, x_2, \dots, x_i, \dots, x_N\}$ , where  $x_i \in X$  is a probable original token of  $y$  and  $\mathbf{c}$  with probability  $\Pr(x_i|y, \mathbf{c})$ . Let the probability that an attacker inferring  $x_i \in X$  as the reconstructed token of  $y$  and  $\mathbf{c}$  be  $p_i$  and it holds that  $\sum_{i=1}^N p_i = 1$ . The expected value of the reconstructed token matching the original token of the sanitized token  $y$  and sanitized context  $\mathbf{c}$  can be expressed as:

$$\Pr(x_1|y, \mathbf{c})p_1 + \Pr(x_2|y, \mathbf{c})p_2 + \dots + \Pr(x_N|y, \mathbf{c})p_N. \quad (19)$$

Utilizing  $\sum_{i=1}^N p_i = 1$ , it can be deduced that:

$$\sum_{i=1}^N \Pr(x_i | y, \mathbf{c})p_i \leq \max_{x' \in X} \Pr(x'|y, \mathbf{c}). \quad (20)$$

According to Equation (20), the expected value of correctly inferring the original token is smaller than  $\max_{x' \in X} \Pr(x'|y, \mathbf{c})$ , which is the expected value under attack strategy in Theorem 2.  $\square$

## B Discussion of Contextual $K$ -ASR Bound

We assess the time required to calculate the Contextual  $K$ -ASR Bound for 1,000 sanitized tokens. we conduct experiments using *CUSTEXT+* (Chen et al., 2023) defense on SST-2 (Socher et al., 2013a) dataset with various values of  $K$ . We set the size of  $X$  in *CUSTEXT+* to its default 20. The experimental results are shown in Table 3. It is observed that as  $K$  increases, the time cost also increases.

On the other hand, the size of the input token set  $X$  in *SANTEXT+* (Yue et al., 2021) far exceeds 20. Specifically, it uses an input token set  $X$  consisting of more than 15,000 tokens in SST-2. Given this, traversing the entire  $X$  for the Contextual  $K$ -ASR Bound in *SANTEXT+* is not practical to implement.

## C Discussion of Context-free Bayesian Attack

The attack strategy in Theorem 1 reconstructs a token considering two factors: a larger  $\Pr(y|x')$  and a larger  $\Pr(x')$ . However, when the probabilities  $\Pr(x')$  of all probable original tokens corresponding to  $y$  are zero in the shadow dataset, the attack strategy in Theorem 1 loses its effectiveness of reconstructing a token with larger  $\Pr(y|x')$ . In such instances, *Context-free Bayesian Attack*, incorporating  $\alpha^{-1}$ , retains its effectiveness. This method selects a reconstructed token  $x'$  according to:

$$x' = \arg \max_{x' \in X} \frac{\Pr(y|x') \cdot (0 + \alpha^{-1})}{\Pr(y)} \quad (21)$$

$$= \arg \max_{x' \in X} \frac{\Pr(y|x') \cdot \alpha^{-1}}{\Pr(y)}. \quad (22)$$

Furthermore,  $\alpha^{-1}$  has minimal impact on the outcome of the reconstruction attack when these probabilities  $\Pr(x')$  are not zero.

To evaluate the practical impact of the parameter  $\alpha^{-1}$ , we conducted an ablation study across three datasets: SST-2 (Socher et al., 2013b), QNLI (Wang et al., 2019b), and AGNEWS (Zhang et al., 2015b). We assessed the Attack Success Rate (ASR) of the *Context-free Bayesian Attack* both with and without  $\alpha^{-1}$  against *SANTEXT+* (Yue et al., 2021), creating 3,000 test samples randomly selected from each dataset.

Figure 2 presents the results of the ablation studies on the parameter  $\alpha^{-1}$ , where the experimental results demonstrate the effectiveness of  $\alpha^{-1}$ .

To further explore the influence of different constants instead of  $\alpha^{-1}$  on *Context-free Bayesian Attack*, we compare the ASR of  $\alpha^{-1}$  with those of other constants against *SANTEXT+*.

Table 4 illustrates the experimental results across three datasets under various privacy parameters  $\epsilon$ . These results highlight the effectiveness of the constant introduced in the *Context-free Bayesian Attack*. When this constant is smaller than  $\alpha^{-1}$ , the ASR of the *Context-free Bayesian Attack* decreases slightly in the experiments. Conversely, when this



Table 3: Time cost of reconstructing 1,000 tokens with various  $K$  in the *Contextual K-ASR Bound*.

$K$	$K = 1$	$K = 5$	$K = 10$	$K = 15$	$K = 20$
Minute	11.631	16.659	18.850	24.566	29.216

Table 4: ASR of *Context-free Bayesian Attack* using various constants instead of  $\alpha^{-1}$  against *SANTEXT+* defense.

Constant	SST-2				QNLI				AGNEWS			
	$\epsilon = 2.0$	$\epsilon = 4.0$	$\epsilon = 2.8$	$\epsilon = 3.2$	$\epsilon = 2.0$	$\epsilon = 4.0$	$\epsilon = 2.8$	$\epsilon = 3.2$	$\epsilon = 2.0$	$\epsilon = 4.0$	$\epsilon = 2.8$	$\epsilon = 3.2$
$\alpha^{-1}$	0.180	0.410	0.650	0.843	0.155	0.434	0.711	0.865	0.223	0.450	0.712	0.894
w/o $\alpha^{-1}$	0.130	0.331	0.600	0.682	0.133	0.387	0.617	0.780	0.204	0.416	0.653	0.820
$0.1 \times \alpha^{-1}$	0.173	0.404	0.648	0.840	0.152	0.429	0.704	0.862	0.218	0.448	0.709	0.892
$5 \times \alpha^{-1}$	0.180	0.410	0.650	0.843	0.155	0.434	0.711	0.865	0.222	0.449	0.711	0.894
$10 \times \alpha^{-1}$	0.180	0.410	0.650	0.843	0.155	0.434	0.711	0.865	0.221	0.449	0.711	0.894

constant exceeds  $\alpha^{-1}$ , the ASR does not show significant growth.

## D Discussion of Contextual Bayesian Attack

To evaluate the effectiveness of the detector  $h$ , we calculate the classification accuracy of both the detector  $h$  and the original BERT model (Devlin et al., 2019) across three datasets: SST-2 (Socher et al., 2013b), QNLI (Wang et al., 2019b), and AGNEWS (Zhang et al., 2015b). We construct 3,000 test samples that are disjointed from the training data of detector  $h$  and randomly selected. For the construction of training samples, we generate 100 sanitized sentences for each original sentence in the shadow dataset. The experiment is conducted 10 times to calculate standard deviations (1-sigma) using the closed form formula (Altman and Bland, 2005) with numpy<sup>1</sup>.

Figure 3 shows the experimental results. It is observed that the classification accuracy of  $h$  is higher than that of the original BERT model. Additionally, the classification accuracy of  $h$  increases with the privacy parameter  $\epsilon$ . These results demonstrate the effectiveness of the detector  $h$  in inferring whether  $c$  is the sanitized context of  $x'$  and  $y$ .

## E Experimental dataset

We conduct the experiments of reconstruction attacks on the following four datasets:

- **SST-2 (Socher et al., 2013b)** : The (binary) Stanford Sentiment Treebank (SST-2) dataset consists of movie reviews with sentiment labels. This dataset is used to categorize the sentiment of sentences (Tian et al., 2020) into

the positive or negative label. We use the version in the GLUE (Wang et al., 2019b).

- **QNLI (Wang et al., 2019b)** : The Question-answering Natural Language Inference (QNLI) dataset is derived from the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016). This dataset consists of question-paragraph pairs, where the task is to determine whether the paragraph contains the answer to the question. We also utilize the use available in the GLUE benchmark (Wang et al., 2019b).
- **AGNEWS (Zhang et al., 2015b)** : The AG News dataset is a widely used benchmark for text classification tasks. It consists of news articles and their corresponding labels. This dataset is utilized to train a model that classifies news into four categories: World, Sports, Business, and Sci/Tech. We use the version of this dataset available on this site<sup>2</sup>.
- **Yelp (Zhang et al., 2015b)** : The Yelp dataset consists of reviews from the Yelp website, where the task is to predict the sentiment expressed in the reviews. The dataset is used for sentiment analysis tasks. We utilized the version available from the Yelp Challenge<sup>3</sup>.

## F Implementation details

### F.1 Context-free ASR Bound and Contextual K-ASR Bound

We run experiments on a cluster with NVIDIA RTX A6000 GPUs and Intel Xeon Gold 6130 2.10 GHz CPUs. We conduct the implementation of all

<sup>1</sup><https://numpy.org/>

<sup>2</sup>AG Corpus of News Articles

<sup>3</sup>Yelp Open Dataset

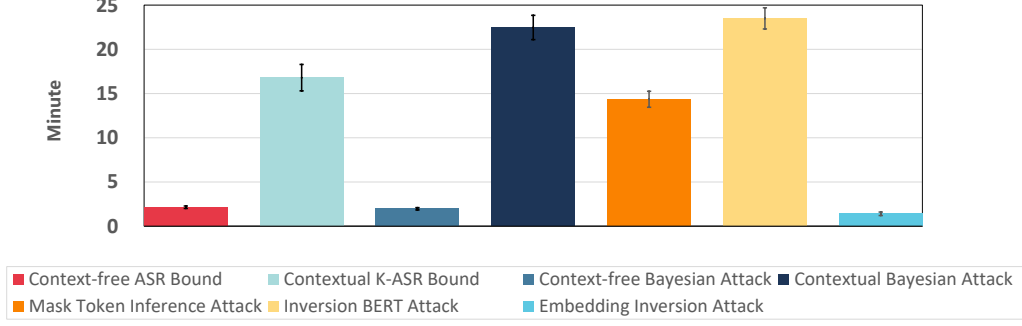


Figure 7: Comparison of the time costs in various reconstruction attacks.

mechanisms using Python. We calculate  $\Pr(x')$  by estimating the statistical probability of the original sentences corresponding to the sanitized tokens.

In the *Contextual K-ASR Bound*, we generate 30 sanitized sentences for each original sentence to construct training samples. For each token  $y \in \Gamma$  in the sanitized sentences, we create a training sample. The input for each sample consists of two token sequences: The first sequence is the sanitized sentence consisting of  $y$  and  $c$ . The second sequence replaces the token  $y \in \Gamma$  in the first sequence with its reconstructed result  $x'$  from *Context-free Bayesian Attack*. If  $x'$  matches the original token of  $y$ , which means that  $c$  is the sanitized context of the sanitized token  $y$  and the reconstructed token  $x'_i$ , we label the sample 1. Otherwise, we label it 0. To maintain label balance, for each sample labeled as 1, we generate a sample labeled as 0 using the same sanitized sentence with the second most likely token from *Context-free Bayesian Attack*. We also generate a sample labeled as 1 with the original token for each sample labeled as 0. For model fine-tuning, we use the pooled output from the BERT model (Devlin et al., 2019) and transform it into class scores through a linear layer.

We use the *Adam Optimizer* (Kingma and Ba, 2015) with a learning rate of 0.00005. We reduce the learning rate by 0.1 every 1000 steps. We measure performance with *CrossEntropyLoss* (Zhang and Sabuncu, 2018) and fine-tune the model over 3 epochs on constructed training samples. The fine-tuned model represents probability  $\Pr(c|x', y)$  that  $c$  is the sanitized context for  $x'$  and  $y$ , based on its output of the label 1.

## F.2 Context-free Bayesian Attack and Contextual Bayesian Attack

We run experiments on a cluster with NVIDIA RTX A6000 GPUs and Intel Xeon Gold 6130 2.10 GHz

CPUs. We conduct the implementation of all mechanisms using Python. We calculate the statistical probability in the shadow dataset to represent the approximated probability  $\Pr(x')$ .

In the *Contextual Bayesian Attack*, we generate 100 sanitized sentences for each original sentence to construct training samples. For each token  $y \in \Gamma$  in the sanitized sentences, we create a training sample. The input for each sample consists of two token sequences: The first sequence is the sanitized sentence consisting of  $y$  and  $c$ . The second sequence replaces the token  $y \in \Gamma$  in the first sequence with its reconstructed result  $x'$  from *Context-free Bayesian Attack*. If  $x'$  matches the original token of  $y$ , which means that  $c$  is the sanitized context of the sanitized token  $y$  and the reconstructed token  $x'$ , we label the sample 1. Otherwise, we label it 0. To maintain label balance, for each sample labeled as 1, we generate a sample labeled as 0 using the same sanitized sentence with the second most likely token from *Context-free Bayesian Attack*. We also generate a sample labeled as 1 with the original token for each sample labeled as 0. For model fine-tuning, we use the pooled output from the BERT model (Devlin et al., 2019) and transform it into class scores through a linear layer.

We use the *Adam Optimizer* (Kingma and Ba, 2015) with a learning rate of 0.00005. We reduce the learning rate by 0.1 every 1000 steps. We measure performance with *CrossEntropyLoss* (Zhang and Sabuncu, 2018) and fine-tune the model over 3 epochs on constructed training samples. The fine-tuned model represents probability  $\Pr(c|x', y)$  that  $c$  is the sanitized context for  $x'$  and  $y$ , based on its output of label 1.

## G Comparison of time cost

To further compare the computation costs, we estimated the time required for various reconstruction

attacks to reconstruct 1,000 sanitized tokens on the SST-2 dataset using the *CUSTEXT+* as a defense. Similarly, we conducted this experiment 10 times to calculate the standard deviations (1-sigma) using the closed form formula (Altman and Bland, 2005) with the numpy<sup>4</sup>.

As illustrated in Figure 7, the time cost for both the *Embedding Inversion Attack* and the *Context-free Bayesian Attack* is approximately comparable, both being less than 5 minutes. These two reconstruction attacks do not require model fine-tuning or inference. In contrast, the *Contextual Bayesian Attack*, *Mask Token Inference Attack*, and *Inversion BERT Attack* involve either model fine-tuning or model inference. Among these three, the *Mask Token Inference Attack* has the smallest time cost, roughly 15 minutes; the *Inversion BERT Attack* takes the longest, about 25 minutes. The time cost of the *Contextual Bayesian Attack* in the experiments is greater than that of the *Mask Token Inference Attack*. This is because *Mask Token Inference Attack* only involves an inference phase without model training. However, our contextual attack requires training a model and inferring on this model. Therefore, *Contextual Bayesian Attack* is slower than the *Mask Token Inference Attack*. Moreover, it is observed that the standard deviations of time costs for the *Embedding Inversion Attack* and the *Context-free Bayesian Attack* are smaller than those for the other reconstruction attacks.

## H Reconstruction attacks with the large epsilon

As illustrated in Figure 4, the performance of our proposed attacks is comparable to that of the *Embedding Inversion Attack* for large values of  $\epsilon$ . The rationale behind this phenomenon is as follows:

Let  $x$  represent a raw token and  $y$  its corresponding sanitized token. The performance of our attacks partially depends on  $\Pr(y|x)$ . Specifically, as  $\epsilon$  increases,  $\Pr(y|x)$  also increases in both *SANTEXT+* and *CUSTEXT+*, causing our attacks to tend to select the token  $x$  with the highest  $\Pr(y|x)$  as the reconstructed token. For a given sanitized token  $y$ , the *Embedding Inversion Attack* selects the token  $x'$  with the smallest embedding distance to  $y$  as the reconstructed token. In the context of *SANTEXT+* and *CUSTEXT+*, this is equivalent to selecting the token  $x'$  with the highest  $\Pr(y|x')$ .

From the above analysis, it can be concluded

that when  $\epsilon$  is large, our proposed attack and the *Embedding Inversion Attack* tend to select the same token as the reconstructed token in most cases.

## I Reconstruction attacks with the misaligned shadow dataset

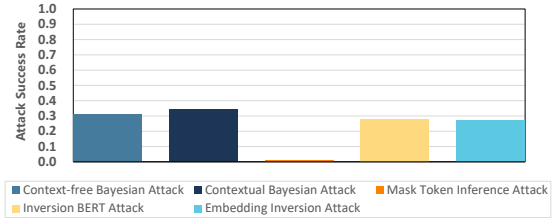


Figure 8: ASR of reconstruction attacks with the misaligned shadow dataset.

In this paper, we present a privacy evaluation study on DP-based text sanitization. Our threat model assumes that practical attacks access a shadow dataset from the same distribution as the target to reconstruct. This assumption simulates the worst-case scenario for privacy preservation and is consistent with the settings described in previous work (Balle et al., 2022; Shokri et al., 2017).

To reflect real-world scenarios where shadow datasets may not be perfectly aligned with sanitized data, we conduct additional attacks on *CUSTEXT+* with  $\epsilon = 3.0$  (as recommended by its authors) using a misaligned shadow dataset. Specifically, we use SST-2 as the shadow dataset and QNLI as the target dataset to reconstruct. Figure 8 presents the experimental results with the misaligned shadow dataset. It is observed that our reconstruction attacks remain comparable to the baseline methods.

<sup>4</sup><https://numpy.org/>