On the Impact of Noise in Differentially Private Text Rewriting

Stephen Meisenbacher, Maulik Chevli, and Florian Matthes
Technical University of Munich
School of Computation, Information and Technology
Department of Computer Science
Garching, Germany
{stephen.meisenbacher,maulikk.chevli,matthes}@tum.de

Abstract

The field of text privatization often leverages the notion of *Differential Privacy* (DP) to provide formal guarantees in the rewriting or obfuscation of sensitive textual data. A common and nearly ubiquitous form of DP application necessitates the addition of calibrated noise to vector representations of text, either at the data- or model-level, which is governed by the privacy parameter ε. However, noise addition almost undoubtedly leads to considerable utility loss, thereby highlighting one major drawback of DP in NLP. In this work, we introduce a new sentence infilling privatization technique, and we use this method to explore the effect of noise in DP text rewriting. We empirically demonstrate that non-DP privatization techniques excel in utility preservation and can find an acceptable empirical privacy-utility trade-off, yet cannot outperform DP methods in empirical privacy protections. Our results highlight the significant impact of noise in current DP rewriting mechanisms, leading to a discussion of the merits and challenges of DP in NLP, as well as the opportunities that non-DP methods present.

1 Introduction

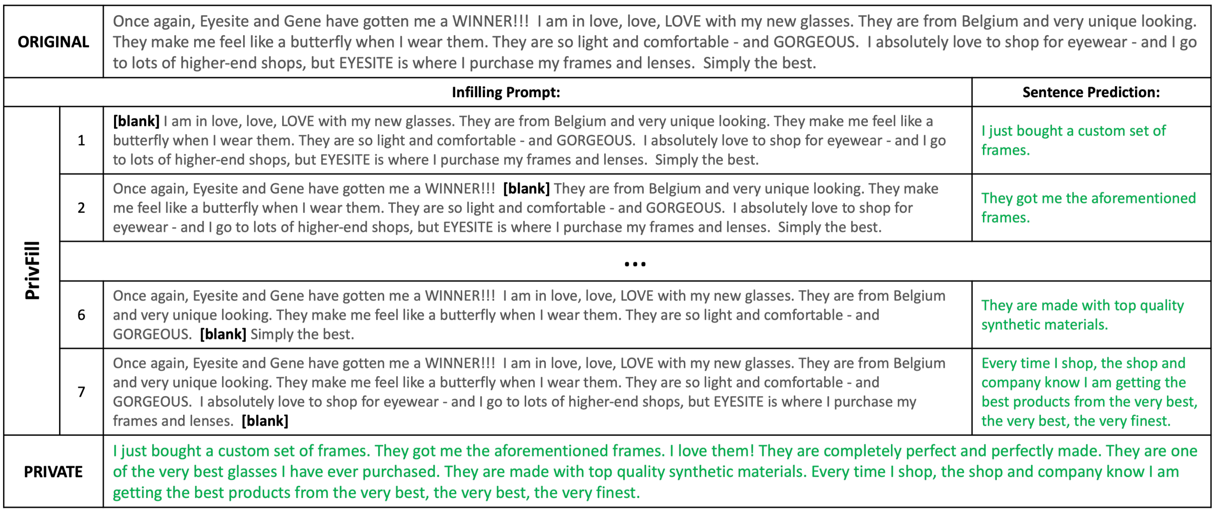
The field of privacy-preserving Natural Language Processing has gained traction in the research community in response to a growing awareness of the need for data privacy in today's AI landscape, particularly with Large Language Models (LLMs) (Pan et al., 2020; Yan et al., 2024). Vulnerabilities of such models have been brought to light by several recent works, mainly in the form of adversarial attacks (Pan et al., 2020; Carlini et al., 2021; Nasr et al., 2023). Beyond the inherent vulnerabilities of the models themselves, other privacy risks emerge as much of the work for NLP applications has moved off local devices to the cloud-based AI service providers (AI-as-a-Service), made most apparent by the boom in usage of ChatGPT and its competitors (Gupta et al., 2023; Wu et al., 2023).

In response to these privacy concerns, numerous solutions have emerged, including the promising and often-utilized privacy-preserving technique of Differential Privacy (DP) (Dwork, 2006; Desfontaines and Pejó, 2020). In essence, DP allows for a quantifiable privacy guarantee, governed by the privacy budget (ε), providing bounded privacy protections for individuals in a dataset. While the study of DP in unstructured domains such as text is less studied and introduces several challenges (Klymenko et al., 2022), numerous technical solutions have been proposed (Hu et al., 2024), from model-level DP to embedding perturbation methods.

Many DP NLP methods model the integration of *local DP* as a *text rewriting* task (Igamberdiev et al., 2022), in which sensitive input texts are rewritten locally to produce "privatized" or "perturbed" output texts before release to third parties. Such a task has been realized in a variety of settings, such as word embedding perturbations (Feyisetan et al., 2020a), latent representation perturbations in encoder-decoder models (Igamberdiev and Habernal, 2023), or DP prompting of LLMs (Utpala et al., 2023). Regardless of where DP is applied, the overarching goal is to produce rewritten texts that are semantically similar to the originals, yet obfuscated enough to mask the direct and indirect information leakage in the originals, thus achieving privacy (Klymenko et al., 2022; Mattern et al., 2022).

While DP in the NLP context offers quantifiable privacy protections, it comes with several challenges and caveats (Meisenbacher and Matthes, 2024b). Among these is clearly reasoning about the syntactic unit protected by DP (Klymenko et al., 2022; Vu et al., 2024), ensuring syntactic and grammatical correctness while privatizing (Mattern et al., 2022), or high-dimensional spaces leading to significantly higher ε values for rewritten outputs to be semantically meaningful (Igamberdiev and Habernal, 2023). Notably, the necessity to add considerable noise to text representations to fulfill

Findings of the Association for Computational Linguistics: NAACL 2025, pages 514–532
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics



ORIGINAL		Once again, Eyesite and Gene have gotten me a WINNER!!! I am in love, love, LOVE with my new glasses. They are from Belgium and very unique looking. They make me feel like a butterfly when I wear them. They are so light and comfortable - and GORGEOUS. I absolutely love to shop for eyewear - and I go to lots of higher-end shops, but EYESITE is where I purchase my frames and lenses. Simply the best.	
		Infilling Prompt:	**Sentence Prediction:**
PrivFill	1	**[blank]** I am in love, love, LOVE with my new glasses. They are from Belgium and very unique looking. They make me feel like a butterfly when I wear them. They are so light and comfortable - and GORGEOUS. I absolutely love to shop for eyewear - and I go to lots of higher-end shops, but EYESITE is where I purchase my frames and lenses. Simply the best.	I just bought a custom set of frames.
	2	Once again, Eyesite and Gene have gotten me a WINNER!!! **[blank]** They are from Belgium and very unique looking. They make me feel like a butterfly when I wear them. They are so light and comfortable - and GORGEOUS. I absolutely love to shop for eyewear - and I go to lots of higher-end shops, but EYESITE is where I purchase my frames and lenses. Simply the best.	They got me the aforementioned frames.
	...		
	6	Once again, Eyesite and Gene have gotten me a WINNER!!! I am in love, love, LOVE with my new glasses. They are from Belgium and very unique looking. They make me feel like a butterfly when I wear them. They are so light and comfortable - and GORGEOUS. **[blank]** Simply the best.	They are made with top quality synthetic materials.
	7	Once again, Eyesite and Gene have gotten me a WINNER!!! I am in love, love, LOVE with my new glasses. They are from Belgium and very unique looking. They make me feel like a butterfly when I wear them. They are so light and comfortable - and GORGEOUS. I absolutely love to shop for eyewear - and I go to lots of higher-end shops, but EYESITE is where I purchase my frames and lenses. **[blank]**	Every time I shop, the shop and company know I am getting the best products from the very best, the very best, the very finest.
PRIVATE		I just bought a custom set of frames. They got me the aforementioned frames. I love them! They are completely perfect and perfectly made. They are one of the very best glasses I have ever purchased. They are made with top quality synthetic materials. Every time I shop, the shop and company know I am getting the best products from the very best, the very best, the very finest.	

Figure 1: PRIVFILL: a sentence infilling text rewriting method that leverages generative LMs to create privatized versions of input documents. For each sentence in an input document, the sentence infilling model is tasked with finding a suitable replacement, and these replacement predictions are concatenated to form a privatized rewritten document. (Pictured: example from the Yelp dataset using PRIVFILL with FLAN-T5-BASE.)

the DP guarantee presents a challenge to achieving practically usable privatization mechanisms. This challenge is exacerbated when privatizing longer text documents (Utpala et al., 2023).

In this work, we investigate these challenges by measuring the effect of noise in DP text rewriting. With this, we pose the following research question:

> *Can language models be leveraged without noise addition for the task of private text rewriting, and what are the empirical utility and privacy results as compared to DP methods with noise?*

Language models have been leveraged for text privation to achieve DP text rewriting in SOTA methods like DP-Prompt (Utpala et al., 2023). These methods, however, utilize LMs to generate privatized texts as a continuation of the original sensitive texts, for example via zero-shot text paraphrasing. This can lead to privatized texts that are considerably shorter than the originals, thereby losing valuable context. To address this, we introduce a new text privatization method based on sentence infilling that allows for greater control and higher preserved context in text rewriting over using a standard prompting technique for text completion.

We introduce a non-DP and DP version of our method, PRIVFILL, allowing for direct comparisons to previously proposed DP mechanisms. We argue that the exhibited ability of our non-DP PRIVFILL variant to act as a usable privatization mechanism makes a compelling case for adoption as opposed to utility-degrading DP methods that must be carefully tuned via ε for desired results. The empirical results, however, showcase the strong privacy-preserving capabilities of mechanisms *with* DP guarantees, providing further empirical evidence of the fine balance between privacy and utility.

The contributions of our work are as follows:

1. We introduce a text privatization mechanism (PRIVFILL), which leverages language model sentence infilling for private text rewriting.
2. We evaluate PRIVFILL, both with and without DP guarantees, in a series of utility and privacy experiments, in comparison to state-of-the-art DP rewriting mechanisms.
3. We showcase the merits of text privatization without noise addition, and we critically discuss the advantages of such an approach in juxtaposition to DP mechanisms.
4. We open-source PRIVFILL, which can be found at: https://github.com/sjmeis/PrivFill

2 Related Work

2.1 Differential Privacy in NLP

Differential Privacy in NLP often leverages the notion of *local* DP (Feyisetan et al., 2020b; Klymenko et al., 2022; Hu et al., 2024; Meisenbacher et al., 2024b), in which text is privatized via some local mechanism, i.e., on the user's device. This is especially applicable for text rewriting, which does not necessitate user texts to be first collected at some central location as would be required for techniques like DP-SGD (Abadi et al., 2016; Kerrigan et al., 2020; Ponomareva et al., 2022).

Text rewriting approaches that integrate DP primarily fall under the category of *embedding perturbation* methods (Hu et al., 2024). Earlier methods first targeted the *word-level* as the unit of protection. Fernandes et al. (2019) and Feyisetan et al. (2020b) proposed the usage of *metric* DP for word-level application, leading to a series of follow-up works that aimed to strengthen the utility preservation of word-level DP mechanisms (Yue et al., 2021; Chen et al., 2023; Carvalho et al., 2023; Meisenbacher et al., 2024a). As pointed out by Mattern et al. (2022), word-level approaches lack contextualization and often struggle with grammatical correctness. As such, other DP NLP works operate directly at the sentence- or document-level (Bo et al., 2021; Weggenmann et al., 2022; Meehan et al., 2022; Igamberdiev and Habernal, 2023). State-of-the-art methods leverage larger language models for DP text rewriting, using the *paraphrasing* task as a proxy (Mattern et al., 2022; Utpala et al., 2023). The main limitation of DP mechanisms targeting document-level text representations, however, is the requirement for high privacy budgets (e.g., 1000) to create meaningful outputs, as highlighted by Igamberdiev and Habernal (2023).

2.2 Non-DP Text Privatization

Beyond DP, stylometric-based methods have been used to replace sensitive words in a text to prevent information leakage of attributes that can be traced to an individual (Reddy and Knight, 2016; Shen et al., 2017). Cyclic machine translation is one of the earliest and still prevalent ways to obfuscate text (Prabhumoye et al., 2018; Xu et al., 2019; Adelani et al., 2021). More recently, style transfer techniques have been used to generate a controlled, obfuscated text using a generative model (Shen et al., 2017; He et al., 2020; Tokpo and Calders, 2022). These privatization methods have not explored the usage of (large) language models; accordingly, we propose a method that addresses this gap, specifically as an alternative to DP methods.

2.3 Text Infilling

Language Models (LMs), particularly recent LLMs, possess strong generative abilities when presented with a text input and tasked to continue or complete the text. The task of *infilling* aims to transform text generation to predict the most likely replacement for some missing span at any point in an input text. Modern text infilling is performed by fine-tuning a generative LM for the task of missing span (sentence or document) prediction. Donahue et al. (2020) propose a simple method for training an LM to fill in any number of missing tokens within a text, and they extend this method to replace entire missing spans. Huang et al. (2020) and Shen et al. (2020) propose similar methods.

Interestingly, Donahue et al. (2020) reports that humans have a difficult time discerning which outputs are created by infillers and which are not. Furthermore, results also indicate that humans sometimes *prefer* rewritten versions over the original (Huang et al., 2020). This motivates the use of infillers for text privatization via a rewriting task, for the purpose of creating coherent, yet distinct and semantically plausible replacements for text documents. To the best of our knowledge, this is the first approach to view the task of text infilling in the same light as private text rewriting. In particular, we focus on sentence infilling, finding the balance between computationally heavy word infilling and potentially incoherent document infilling.

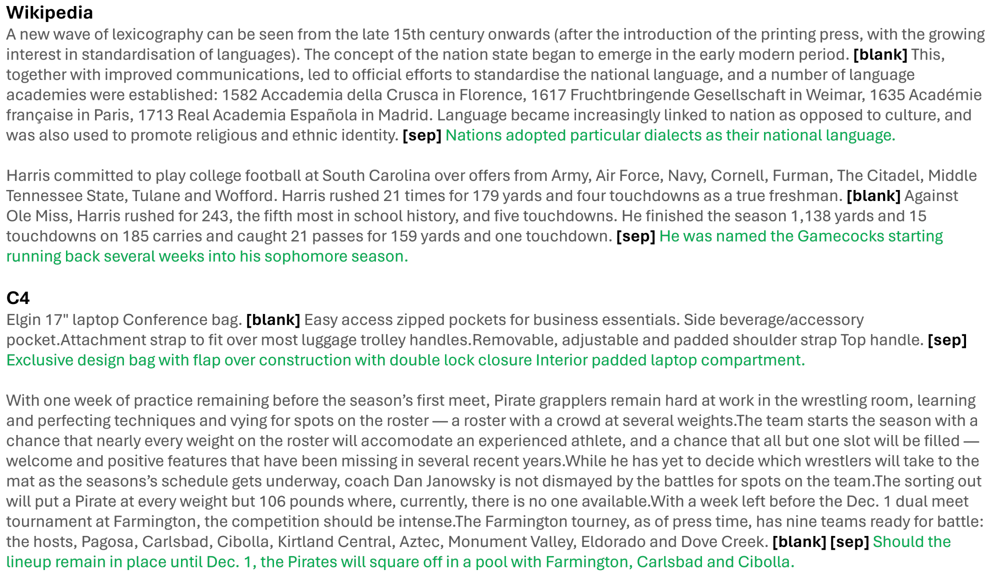
3 A Simple Infilling-based Text Privatization Mechanism

We introduce PrivFill, a simple text rewriting mechanism that rewrites text documents by leveraging fine-tuned LM infilling models. We outline the process undertaken for infiller training, as well as describe the underlying mechanism of PrivFill.

3.1 Training a Sentence Infiller

To train our base sentence infilling models for PrivFill, we follow the approach proposed by Donahue et al. (2020), slightly modified for our objectives. To create training samples for sentence infilling, we employ two datasets: English Wikipedia and Common Crawl. The exact dataset preparation details for each dataset are described next.

Wikipedia. We select the train split of the `20220301.en` version of the English Wikipedia dump (Wikimedia-Foundation, 2022). Next, we randomly sample one million articles, and from each, we randomly select one sentence to mask out. Due to the potentially large size of articles, we only keep the preceding two and ensuing two context sentences, if available. Thus, the resulting infilling dataset contains one million samples where the missing sentence is “centered” in context.



Wikipedia

A new wave of lexicography can be seen from the late 15th century onwards (after the introduction of the printing press, with the growing interest in standardisation of languages). The concept of the nation state began to emerge in the early modern period. **[blank]** This, together with improved communications, led to official efforts to standardise the national language, and a number of language academies were established: 1582 Accademia della Crusca in Florence, 1617 Fruchtbringende Gesellschaft in Weimar, 1635 Académie française in Paris, 1713 Real Academia Española in Madrid. Language became increasingly linked to nation as opposed to culture, and was also used to promote religious and ethnic identity. **[sep]** Nations adopted particular dialects as their national language.

Harris committed to play college football at South Carolina over offers from Army, Air Force, Navy, Cornell, Furman, The Citadel, Middle Tennessee State, Tulane and Wofford. Harris rushed 21 times for 179 yards and four touchdowns as a true freshman. **[blank]** Against Ole Miss, Harris rushed for 243, the fifth most in school history, and five touchdowns. He finished the season 1,138 yards and 15 touchdowns on 185 carries and caught 21 passes for 159 yards and one touchdown. **[sep]** He was named the Gamecocks starting running back several weeks into his sophomore season.

C4

Elgin 17" laptop Conference bag. **[blank]** Easy access zipped pockets for business essentials. Side beverage/accessory pocket.Attachment strap to fit over most luggage trolley handles.Removable, adjustable and padded shoulder strap Top handle. **[sep]** Exclusive design bag with flap over construction with double lock closure Interior padded laptop compartment.

With one week of practice remaining before the season’s first meet, Pirate grapplers remain hard at work in the wrestling room, learning and perfecting techniques and vying for spots on the roster — a roster with a crowd at several weights.The team starts the season with a chance that nearly every weight on the roster will accomodate an experienced athlete, and a chance that all but one slot will be filled — welcome and positive features that have been missing in several recent years.While he has yet to decide which wrestlers will take to the mat as the seasons’s schedule gets underway, coach Dan Janowsky is not dismayed by the battles for spots on the team.The sorting out will put a Pirate at every weight but 106 pounds where, currently, there is no one available.With a week left before the Dec. 1 dual meet tournament at Farmington, the competition should be intense.The Farmington tourney, as of press time, has nine teams ready for battle: the hosts, Pagosa, Carlsbad, Cibolla, Kirtland Central, Aztec, Monument Valley, Eldorado and Dove Creek. **[blank]** **[sep]** Should the lineup remain in place until Dec. 1, the Pirates will square off in a pool with Farmington, Carlsbad and Cibolla.

Figure 2: Samples from our infilling train dataset (target infilling text in green, **[sep]** token inserted for readability).

Common Crawl (C4). We employ the *Colossal Clean Crawled Corpus* (C4) made available by Raffel et al. (2020), using the en (cleaned) split. We randomly sample one million rows, and within each sampled text, we randomly select one sentence to mask. As opposed to Wikipedia, we preserve the entire text, due to the smaller average text length.

Final Dataset. The final infilling train dataset is created by merging the two abovementioned prepared datasets. As done by Donahue et al. (2020), we utilize the `[blank]` token to represent the masked-out sentences. We do not use the `[answer]` token to demarcate the target text spans, as we only mask one sentence per training instance. Examples originating from both training corpora can be found in Figure 2.

3.1.1 Model Training

We choose three base model checkpoints for fine-tuning on our infilling dataset: BART-LARGE (Lewis et al., 2019), FLAN-T5-BASE, and FLAN-T5-LARGE (Chung et al., 2022). These models were chosen for two reasons: (1) they are powerful generative models that are still (relatively) small in size, thus not too resource-intensive, and (2) they present an opportunity to compare with selected DP methods, discussed in Section 4.1.

For fine-tuning, we use the HuggingFace Trainer. All models were trained for one epoch with a learning rate of 5e-5 on a 99% training split of our infilling dataset. The ROUGE metric (Lin, 2004) was used for evaluation. All other default training parameters were kept. For performance, we limit all models to a max length of 512 tokens.

3.2 PRIVFILL

We now introduce our PRIVFILL mechanism, a simple infilling privatization mechanism that rewrites text documents sentence-by-sentence.

Algorithm 1

PRIVFILL Text Rewriting Mechanism

```
Require: Infilling Model $\mathcal{M}$, Tokenizer $\mathcal{T}$, input document $d$, max length $max\_len$, max new tokens $mnt$
Ensure: privatized document $priv$
  sentences ← $sent\_tokenize(d)$
  rewritten ← [ ]
  for $idx \in 1...len($sentences$)$ do
      prompt ← $d.replace($sentences$[idx]$, “[blank]”$)$
      inputs ← $\mathcal{T}.encode($prompt$, max\_len)$
      output ← $\mathcal{M}.generate($inputs$, mnt, do\_sample)$
      decoded ← $\mathcal{T}.decode($output$)$
      rewritten.$append($decoded$)$
  end for
  $priv = concatenate($rewritten$)$
  return $priv$
```

As specified in Algorithm 1, the PRIVFILL mechanism rewrites a document in sequential order of the sentences from the input document. The privatized output document is thus comprised of the predicted sentences from the infilling model for each of the component sentences. In this way, we can create privatized documents composed of contextualized sentence replacements. Due to the nature of infilling, PRIVFILL would require documents with at least two sentences for sensible outputs, although this is not a necessary condition.

Algorithm 2
PrivFillDP Text Rewriting Mechanism

```
Require: Infilling Model $\mathcal{M}$, Tokenizer $\mathcal{T}$, input document $d$, max length $max\_len$, max new tokens $mnt$, privacy budget $\varepsilon$, sensitivity $\Delta$
Ensure: privatized document $priv$
  temperature $\leftarrow \frac{2 \cdot \Delta}{\varepsilon}$
  sentences $\leftarrow sent\_tokenize(d)$
  rewritten $\leftarrow [\,]$
  for $idx \in 1...len($sentences$)$ do
      prompt $\leftarrow d.replace($sentences$[idx]$, “[blank]$'')$
      inputs $\leftarrow \mathcal{T}.encode($prompt$, max\_len)$
      logits $\leftarrow \mathcal{M}.encode($inputs$)$
      logits $\leftarrow$ logits/temperature
      output $\leftarrow \mathcal{M}.generate($logits$, mnt, do\_sample)$
      decoded $\leftarrow \mathcal{T}.decode($output$)$
      rewritten$.append($decoded$)$
  end for
  $priv = concatenate($rewritten$)$
  return $priv$
```

3.3 PrivFillDP

We also construct a variant of PrivFill that fulfills DP guarantees, in order to evaluate the effect of DP (i.e., noise). To transform PrivFill into PrivFillDP, we utilize a DP token selection mechanism based on the Exponential Mechanism (Mattern et al., 2022; Utpala et al., 2023; Meisenbacher et al., 2024a). Specifically, we use the fine-tuned flan-t5-large infilling model described previously, and in the generation phase, tokens are sampled in the same way as with DP-Prompt (see Section 4.1); hence, the theoretical DP guarantees are the same as DP-Prompt. We use the same clipping bounds and ε values as reported in Section 4.1, to ensure direct comparability to a SOTA DP mechanism. For further details on the DP token selection mechanism based on temperature sampling, we refer the reader to the original works mentioned above and to Algorithm 2, which outlines the pseudocode for PrivFillDP.

4 Experimental Setup

We describe our experimental design, specifically the tasks assigned to both utility and privacy experiments. These experiments evaluate the utility- and privacy-preserving capabilities of PrivFill comparatively to previous DP rewriting methods.

4.1 Selected DP Mechanisms

DP-BART (Igamberdiev and Habernal, 2023). DP-BART introduces a DP mechanism to the latent representation between the encoder and decoder blocks of a BART model (Lewis et al., 2019). In essence, a noisy encoder representation is fed to the BART decoder, which then outputs a privatized version of the original text. We test the **DP-BART-CLV** version as proposed by Igamberdiev and Habernal (2023), as this does not require further training. For ε values, we choose $\varepsilon \in \{500, 1000\}$, as these roughly represent the median and third quartile of evaluated values in the original paper.

DP-Prompt (Utpala et al., 2023). DP-Prompt leverages zero-shot prompting of LLMs to rewrite texts by *paraphrasing*. In particular, DP is ensured by the utilization of a temperature sampling mechanism that mimics the behavior of the well-known Exponential Mechanism (McSherry and Talwar, 2007). The authors of DP-Prompt test their mechanism on a variety of open- and closed-source LLMs. For our evaluation, we utilize Flan-T5-Large (Chung et al., 2022) due to it being open-source, not too resource intensive (780M parameters), and directly comparable to our PrivFill base models.

For DP-Prompt, the calculation and choice of ε value depends on the clipping range chosen for the model logit values. Following Utpala et al. (2023), we measure the observed range of logit values for our chosen Flan-T5-Large model, run on 100 randomly selected texts from the C4 corpus, and accordingly choose the clipping range to be $(logit_{min}, logit_{max}) = (-95, 8)$, which is normalized to the range of $[0, 1]$[1], thus leading to a sensitivity of 1. Next, we choose the temperature values T of 1.0 and 1.5, representing the first and third quartile values used in the original implementation[2]. Under the temperature sampling mechanism, this equates to values of $\varepsilon \in \{1.\overline{3}, 2\}$[3].

4.2 Utility Experiments

We employ three datasets to test the utility preservation of our selected mechanisms on document classification tasks. The results of these experiments are found in Table 1.

4.2.1 Datasets and Tasks

arXiv Abstracts. The *arXiv Paper Abstracts* dataset[4] is a collection of nearly 40,000 arXiv papers, specifically the title and abstract. Each paper is classified into one or more arXiv paper categories.

[1] as done by Mattern et al. (2022)

[2] As only five values were tested by Utpala et al. (2023), we opt for different quartile values than with DP-BART.

[3] Following the formula $\varepsilon = \frac{2 \cdot \Delta}{T}$, where Δ represents sensitivity. Here, ε is *per generated token*.

[4] https://www.kaggle.com/datasets/spsayakpaul/arxiv-paper-abstracts

Task	arXiv (10 classes)					BBC (5 classes)					DocNLI (2 classes)				
Baseline	$46.58_{1.0}$ (42)					$98.52_{0.3}$ (24)					$86.87_{1.8}$ (48)				
Mechanism	R1	RL	CS	PPL	F1 ↑	R1	RL	CS	PPL	F1 ↑	R1	RL	CS	PPL	F1 ↑
DP-BART ($\varepsilon = 500$ / document)	0.21	0.14	0.08	26	$29.53_{0.0}$	0.25	0.14	0.17	24	$46.67_{1.3}$	0.21	0.14	0.18	29	$75.10_{2.5}$
DP-BART ($\varepsilon = 1000$ / document)	0.30	0.18	0.22	11	$30.81_{1.8}$	0.36	0.17	0.45	11	$89.42_{1.0}$	0.34	0.21	0.47	12	$\mathbf{76.48_{0.5}}$
DP-Prompt ($\varepsilon = 1.\overline{3}$ / token)	0.05	0.04	0.17	8773	$29.53_{0.0}$	0.04	0.02	0.20	9895	$29.31_{1.9}$	0.05	0.03	0.20	10762	$54.34_{3.4}$
DP-Prompt ($\varepsilon = 2$ / token)	0.11	0.08	0.58	612	$34.78_{1.0}$	0.07	0.06	0.59	809	$88.15_{0.6}$	0.11	0.09	0.48	1283	$60.14_{3.5}$
PrivFillDP ($\varepsilon = 1.\overline{3}$ / token)	0.13	0.06	0.20	7945	$31.48_{1.4}$	0.18	0.07	0.35	6856	$89.95_{0.5}$	0.12	0.06	0.25	9376	$69.47_{2.9}$
PrivFillDP ($\varepsilon = 2$ / token)	0.32	0.14	0.53	468	$34.69_{0.3}$	0.39	0.14	0.65	309	$96.61_{0.4}$	0.29	0.13	0.52	496	$71.66_{0.6}$
PrivFill (bart-large)	0.55	0.28	0.83	15	$40.10_{4.2}$	0.54	0.25	0.82	9	$97.04_{0.4}$	0.47	0.26	0.75	15	$73.71_{1.8}$
PrivFill (flan-t5-base)	0.47	0.21	0.76	48	$39.12_{4.2}$	0.51	0.19	0.80	31	$97.99_{0.7}$	0.40	0.18	0.69	46	$74.51_{2.5}$
PrivFill (flan-t5-large)	0.48	0.21	0.77	45	$\mathbf{42.18_{2.9}}$	0.51	0.19	0.80	27	$\mathbf{98.20_{0.4}}$	0.41	0.18	0.70	40	$73.41_{0.7}$

Table 1: Utility Results. *R1*, *RL*, *CS*, and *PPL* denote ROUGE-1/-L, cosine similarity, and perplexity, respectively. The highest F1 for each task is **bolded**. Subscripts denote standard deviations. Baseline PPL is given in parentheses.

We only consider papers belonging to the top 10 most frequent categories, namely {*cs.AI*, *cs.LG*, *stat.ML*, *cs.CV*, *cs.CL*, *eess.IV*, *cs.RO*, *cs.CR*, *cs.NE*, *cs.GR*}. Papers with multiple labels are considered only by their top (first) label. This results in a final dataset of 26,109 rows. With this dataset, we create a 10-class single-label classification task.

BBC News. The *BBC News* dataset[5] is a corpus of 3147 BBC News articles from five categories: {*business*, *entertainment*, *politics*, *sports*, *tech*}, creating a 5-class single-label classification task.

DocNLI. The *DocNLI* dataset (Yin et al., 2021) was created to introduce a *document*-level entailment prediction task, as opposed to previous approaches focusing on sentences. The dataset consists of (*premise*, *hypothesis*) pairs, each marked as *entailment* or *not entailment*. Since the original dataset is expansive, we take a 1% random sample, resulting in a 9136-row dataset. The dataset is then used for a two-class binary classification task.

4.2.2 Evaluation Strategy

For each of the three datasets, we first achieve baseline (micro) F1 Scores by fine-tuning a DEBERTA-V3-BASE (He et al., 2021) model on the original datasets. Each dataset is separated into a 90/10 split. The model is trained on the train split for one epoch using a learning rate of 5e-5 and otherwise default HuggingFace TRAINER parameters. Training is repeated three times with a different shuffled version of the train split. All reported metrics represent the trained model's performance on the validation split, averaged over the three evaluations.

Next, each of the privatization approaches is used to rewrite the datasets. These rewritten datasets are then used for the same fine-tuning procedure, and the final evaluation metrics are captured for comparison to the baseline. Thus, the experiments measure the ability of each rewriting approach to preserve the utility of the original data.

4.2.3 Additional Metrics

In addition to F1, we report three additional metrics usually employed in text generation evaluation: ROUGE (Lin, 2004) (i.e., ROUGE-1/-L), cosine similarity (CS, inspired by BERTScore (Zhang et al., 2020), and perplexity (PPL). While not explicit measures of utility preservation, these three metrics are useful in explaining the relation between the original and rewritten texts, e.g., whether they are similar in (n-gram) token overlap and/or semantic similarity. As such, the scores will be useful to the analysis of the results presented later in this section. Note that for the CS score, we use the ALL-MINILM-L6-V2 model from SENTENCE-TRANSFORMERS. PPL is used to measure the *plausibility* and *naturalness* of the privatized texts (Weggenmann et al., 2022; Meisenbacher and Matthes, 2024a). We report mean PPL using a GPT-2 model. (Radford et al., 2019).

4.3 Privacy Experiments

For the empirical measurement of privacy preservation of PRIVFILL(DP) in comparison to the selected DP mechanisms, we perform a two-part privacy experiment. The results of these experiments are found in Table 2.

4.3.1 Empirical Privacy: Datasets and Tasks

To test the privacy preservation offered by a text rewriting mechanism, evaluation approaches often test *empirical privacy*, where the goal is to reduce the ability of an adversary to infer sensitive information given a text. We employ three datasets for empirical privacy testing for two adversarial settings: gender and authorship identification.

Trustpilot Reviews. The *Trustpilot Reviews* corpus (Hovy et al., 2015) is a large-scale collection

[5] http://mlg.ucd.ie/datasets/bbc.html

Trustpilot (Gender)	Baseline	DP-BART ($\varepsilon = 500$)	DP-BART ($\varepsilon = 1000$)	DP-Prompt ($\varepsilon = 1.\overline{3}$)	DP-Prompt ($\varepsilon = 2$)	PrivFillDP ($\varepsilon = 1.\overline{3}$)	PrivFillDP ($\varepsilon = 2$)	PrivFill (BART-LARGE)	PrivFill (FLAN-T5-BASE)	PrivFill (FLAN-T5-LARGE)
Utility F1 ↑	$99.57_{0.1}$	$93.59_{0.2}$	$98.16_{0.0}$	$92.18_{0.3}$	$95.13_{0.1}$	$93.53_{1.2}$	$96.80_{0.1}$	$98.63_{0.1}$	$97.33_{0.2}$	$98.73_{0.1}$
PP+↑	289	-309	148	-450	-155	-315	12	195	65	205
Privacy F1 (static) ↓	72.46	59.61	59.47	58.56	62.16	58.12	62.26	60.16	61.00	60.73
Privacy F1 (adapt.) ↓	72.46	$58.09_{0.0}$	$60.38_{1.7}$	$58.09_{0.0}$	$60.10_{1.5}$	$58.09_{0.0}$	$59.36_{1.8}$	$65.54_{0.3}$	$61.53_{2.6}$	$62.73_{0.8}$
Relative Gain (static) ↑	-	0.62	**1.87**	0.42	0.59	0.86	1.02	**1.87**	1.38	1.80
Relative Gain (adapt.) ↑	-	0.88	**1.71**	0.51	0.95	0.87	1.52	0.94	1.29	1.46

Yelp (Author)	Baseline	DP-BART ($\varepsilon = 500$)	DP-BART ($\varepsilon = 1000$)	DP-Prompt ($\varepsilon = 1.\overline{3}$)	DP-Prompt ($\varepsilon = 2$)	PrivFillDP ($\varepsilon = 1.\overline{3}$)	PrivFillDP ($\varepsilon = 2$)	PrivFill (BART-LARGE)	PrivFill (FLAN-T5-BASE)	PrivFill (FLAN-T5-LARGE)
Utility F1 ↑	$95.03_{0.1}$	$93.52_{0.0}$	$94.30_{0.5}$	$93.52_{0.0}$	$93.49_{0.0}$	$93.53_{0.0}$	$93.60_{0.3}$	$94.32_{0.6}$	$94.37_{0.1}$	$93.95_{0.5}$
PP+↑	-132	-283	-205	-283	-286	-282	-275	-203	-198	-240
Privacy F1 (static) ↓	96.30	21.85	16.99	17.05	20.17	25.26	43.01	24.86	40.11	50.00
Privacy F1 (adapt.) ↓	96.30	$42.16_{0.9}$	$61.14_{0.9}$	$18.98_{0.2}$	$19.44_{1.5}$	$35.84_{0.9}$	$61.02_{0.9}$	$70.77_{0.6}$	$62.25_{0.8}$	$67.09_{1.1}$
Relative Gain (static) ↑	-	2.05	1.64	**2.13**	2.10	1.06	1.69	1.51	1.25	1.36
Relative Gain (adapt.) ↑	-	1.74	0.98	2.10	**2.11**	1.84	1.41	0.82	0.92	1.10

Enron Emails (Author)	Baseline	DP-BART ($\varepsilon = 500$)	DP-BART ($\varepsilon = 1000$)	DP-Prompt ($\varepsilon = 1.\overline{3}$)	DP-Prompt ($\varepsilon = 2$)	PrivFillDP ($\varepsilon = 1.\overline{3}$)	PrivFillDP ($\varepsilon = 2$)	PrivFill (BART-LARGE)	PrivFill (FLAN-T5-BASE)	PrivFill (FLAN-T5-LARGE)
Privacy F1 (static) ↓	45.89	13.61	23.05	6.93	16.48	2.77	2.60	18.14	15.62	15.22
Privacy F1 (adapt.) ↓	45.89	$11.26_{0.9}$	$23.38_{0.4}$	$8.46_{0.9}$	$14.94_{0.4}$	$8.22_{0.6}$	$11.53_{0.5}$	$18.28_{1.3}$	$15.76_{1.2}$	$18.25_{1.0}$

Table 2: Empirical Privacy Results. Subscripts denote standard deviations. Note that *relative gain* is not reported for Enron Emails (no associated utility task). *PP+* indicates the percentage points above/below majority class guessing.

of reviews from the Trustpilot platform. Each review is assigned a review score (1-5), as well as the gender (M/F) of the reviewer. We take a 10% random sample of reviews containing two or more sentences, and create a binary classification task from negative reviews (1-2) and positive reviews (5). This results in a dataset of 29,490 rows.

Yelp Reviews. We utilize a subset of the *Yelp Reviews* dataset, as used by Utpala et al. (2023) in their privacy experiments. The dataset consists of 17,295 reviews from 10 distinct users (denoted by user ID), each marked as positive or negative.

Enron Emails. The *Enron Email Dataset*[6] consists of over 600,000 emails from 158 employees of the Enron Corporation, shortly before its collapse in 2001. The corpus has been used for a plethora of NLP explorations, but in this work, we use emails from the *sent_items* folders of each email user for an authorship identification task. To prepare the dataset, we use emails only containing 2 or more sentences, and we only use emails from users having written the most emails[7], resulting in a dataset of 28 users' sent emails. Emails were cleaned to remove headers and obvious identifiers, a process that is outlined in more detail in Appendix A. This resulted in a final dataset of 12,283 emails.

4.3.2 Empirical Privacy: Evaluation

The Trustpilot and Yelp datasets present the opportunity for a two-sided evaluation of utility and privacy. First, we fine-tune a DEBERTA-V3-BASE model for binary sentiment classification (measured by F1), similarly to the utility experiments.

For empirical privacy, we model two types of attackers: *static* and *adaptive* (Mattern et al., 2022; Utpala et al., 2023). In the static setting, an adversarial model (i.e., classification of gender or author) is trained only on the original texts, and the empirical privacy is measured by F1 score of the model on rewritten datasets. In the adaptive setting, the model is trained *and* evaluated on the rewritten texts, thus modeling a more formidable adversary who has knowledge of the exact rewriting process. In both cases, training is done for one epoch.

With Trustpilot and Yelp, we also report the *relative gain* metric (Weggenmann et al., 2022; Mattern et al., 2022). This metric captures the relative advantage of rewriting text (privacy gain) over the potential utility loss. Let P_o, U_o represent the baseline privacy and utility scores, respectively, and P_r, U_r be the scores observed on the privatized datasets. The relative gain is thus defined as $RG = (U_r/U_o) - (P_r/P_o)$, with the higher the better. Note that we calculate the change in F1 over random / majority-class guessing on the validation set, denoted MG_u (utility) and MG_p (privacy), as the two datasets are imbalanced; thus $RG = \frac{U_r - MG_u}{U_o - MG_u} - \frac{P_r - MG_p}{P_o - MG_p}$. Please refer to Appendix B for more details on MG_u and MG_p.

5 Discussion

The Impact of Noise. In comparing the results of PrivFill to PrivFillDP, one can immediately observe the effect of the noise as a result of the DP token selection mechanism. On all tasks, the non-

[6] https://www.cs.cmu.edu/~enron/

[7] Specifically, 388 or more emails (80th percentile).

DP version performs significantly better in F1, CS, ROUGE, and PPL scores, even when compared to the weaker $\varepsilon = 2$ variant. Beyond the captured metrics, the impact of noise becomes apparent when viewing the example outputs of Appendix B, which showcase the fluent texts of PrivFill as opposed to the lower-quality texts of the DP version.

The merits of noise addition to achieve DP must also be highlighted. In the privacy experiments, DP methods are generally more effective in reducing adversarial advantage, particularly showcased in the Yelp task. Comparing PrivFill to PrivFillDP, the results are similar against the static attacker, but the non-DP privacy protections begin to be weakened in the adaptive setting.

Taking into account the utility loss and relative gains, the results provide some interesting lessons. While PrivFill almost always leads to less utility loss, as exhibited in Tables 1 and 2, this advantage does not always lead to higher relative gains due to the lower empirical privacy protection. Interestingly, relative gain also depends on the privacy task at hand and the modeled adversary; for example, PrivFill performs better against DP methods in the Trustpilot task than the Yelp task.

Our results also uncover the importance of rewriting mechanism design. Although both DP-Prompt and PrivFillDP utilize the same underlying "noise" mechanism, i.e., DP token selection via temperature, PrivFillDP nearly always outperforms DP-Prompt on all measured utility metrics when comparing the same ε values. This can plausibly be attributed to the infilling mechanism of PrivFill versus the reliance on zero-shot LLM paraphrase prompting in DP-Prompt. Thus, we empirically demonstrate that the impact of noise in DP text rewriting is greatly influenced by the design of the rewriting mechanism itself.

Utility *and* Privacy <u>without</u> Noise? The empirical results demonstrated by PrivFill show that the generative power of (L)LMs can be effectively leveraged for the task of private text rewriting. We see that PrivFill exhibits particular strength in rewriting *topical* documents, such as those in arXiv and BBC (Table 1). Interestingly, only PrivFill is able to maintain high PPL scores while also preserving semantic similarity (in CS). In contrast, DP-BART scores well on PPL, but with very low CS, while DP-Prompt scores poorly in both.

The empirical results show the fine balance between a privacy-preserving text rewriting mechanism, and one that also preserves utility. Table 2 exhibits that although PrivFill is validated in its ability to reduce adversarial advantage, the magnitude of such a decrease is often less than that of the DP mechanisms. However, this comes at the trade-off of higher utility preservation by PrivFill. We hold that these results illustrate a balance between privacy and utility, yet the exact nature of evaluating the trade-off is an open research area.

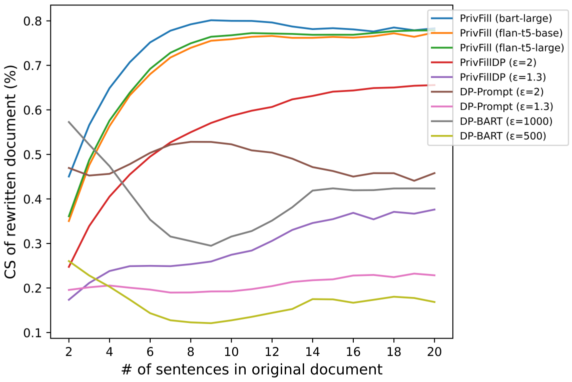


Figure 3: Cosine Similarity (CS) of rewritten texts vs. number of sentences in the original document.

Beyond metrics, the impact of noise can also be demonstrated qualitatively. Looking at example outputs of all tested mechanisms (see Appendix B), one can observe the strengths of our method in the readability, coherence, and plausibility of rewritten outputs. This comes in juxtaposition to DP rewritten texts, which often suffer from nonsensical, repetitive, off-topic, or unnecessarily short (in the case of paraphrasing) outputs, even at high ε values. The requirement for meaningful privatized text would certainly be crucial to the practical adoption of private text rewriting.

Revisiting DP. In evaluating private text rewriting at the *document*-level, our findings reveal particular difficulties experienced by DP rewriting methods, namely in rewriting longer texts. In Figure 3, we analyze the performance of all methods (in CS) based upon input document sentence count. As can be observed, while PrivFill becomes stronger with longer documents (i.e., more context), DP methods operate on a plateau. This is alleviated to a degree with PrivFillDP, which does not perform as well, but still follows a similar curve.

These findings open a discussion regarding the merits of current DP text rewriting. While certain advantages are clear, such as strong empirical privacy results (Table 2), one cannot ignore the significant effect that achieving DP entails. Even at high

ε values, all DP methods achieve poor ROUGE and CS scores (Table 1), a deficiency the non-DP PRIVFILL rectifies. Harking back to the previous point about the significance of mechanism design, our findings also show that when designing text privatization methods, DP or not, it becomes important to consider the nature of the target texts to be privatized. Therefore, latent space noise (DP-BART) or zero-shot paraphrasing (DP-PROMPT) might not be as effective for (long) document privatization.

With such implications, one must critically view the meaning of ε in DP rewriting settings. Taking the example of DP-BART, we see that even at large ε values, the generated text is both syntactically and semantically distant from the input texts. In practice, one may try to improve this by using DP-BART on the sentence level and subsequently composing the privacy budgets, similar to PRIVFILLDP. This, however, would only exacerbate the difficulty in interpreting ε and force the total privacy budgets to increase significantly. We argue, therefore, that in light of our observed results, striving for such weak guarantees only serves to achieve DP in name, but this directly affects utility that could be regained without DP.

The questions raised here call for a discussion on the practical advantage that non-DP brings to text rewriting. We find that despite the lack of a mathematically grounded privacy guarantee, PRIVFILL demonstrates comparable (albeit, slightly worse) empirical privacy protection to its DP counterpart, yet with the benefit of maintaining higher semantic coherence (CS). In addition, it offers competitive, and in some cases, better relative gains than DP methods. At the same time, only DP methods, particularly those at lower ε values, are able to neutralize adversarial advantage to majority- or random-class guessing, particularly in the adaptive attacker setting, something which the non-DP PRIVFILL cannot offer. This, as stated, comes at the cost of likewise diminished data utility.

6 Conclusion

In this work, we introduce PRIVFILL as an alternative to DP rewriting requiring utility-degrading noise addition. We evaluate PRIVFILL, both with and without DP guarantees, in utility and privacy experiments, which shed light on the significant impact of noise addition on utility while also highlighting the ability of noise to provide strong privacy protections. We discuss the merits of DP versus non-DP text privatization, calling for more research at this important intersection. Above all, we advocate that future research in DP NLP should focus on usability and critical analyses of the practical advantages that theoretical guarantees offer.

The discussion in our work gives way to suggestions for future research to advance the field of privacy-preserving NLP, namely further explorations into (1) non-DP text privatization, (2) new DP techniques that maintain meaningful privacy budgets, and accordingly, (3) improving the explainability of DP in NLP, which is crucial to moving the field forward from research to practice.

Limitations

A limitation of our work is the sole focus on *sentence* infilling in PRIVFILL; we do not experiment with token and/or document infilling. Although we hypothesize that sentence infilling is most apt for the task, follow-up work should include a detailed analysis of different infilling granularities.

Another limitation is rooted in the focus on DP vs. non-DP, where we position our proposed method directly in juxtaposition to SOTA DP methods. We do not, however, evaluate against other text privatization methods outside of the DP realm. Future work should further validate the efficacy of PRIVFILL(DP) against these additional methods.

Ethics Statement

While the Trustpilot, Yelp, and Enron datasets were not originally intended for use in adversarial scenarios, we do not inflict any harm as these datasets are already publicly available. Furthermore, the Trustpilot and Yelp datasets contain no personally identifiable information (PII). Although the Enron dataset does contain PII, the dataset was made public by the Federal Energy Regulatory Commission during its investigation of the company post-2001.

Another ethical consideration is rooted in the known dangers of using generative AI, namely LLMs. Specifically, LLMs are known to hallucinate in their use for generation, sometimes resulting in seemingly plausible but ultimately untrue or misinformed outputs. In this work, we do not address this issue, and in our PRIVFILL mechanism, we make no attempt to verify the information contained in the privatized outputs, as this is not an objective of optimal text privatization. Nevertheless, caution should be exercised in the potential use and release of data stemming from our method.

Acknowledgments

The authors thank the anonymous reviewers for their feedback, as well as Alexandra Klymenko for her valuable contributions to this work.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 308–318, New York, NY, USA. Association for Computing Machinery.

David Adelani, Miaoran Zhang, Xiaoyu Shen, Ali Davody, Thomas Kleinbauer, and Dietrich Klakow. 2021. Preventing author profiling through zero-shot multilingual back-translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8687–8695, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haohan Bo, Steven H. H. Ding, Benjamin C. M. Fung, and Farkhund Iqbal. 2021. 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*, pages 3997–4007, Online. Association for Computational Linguistics.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Ricardo Silva Carvalho, Theodore Vasiloudis, Oluwaseyi Feyisetan, and Ke Wang. 2023. TEM: High utility metric differential privacy on text. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pages 883–890. SIAM.

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, Toronto, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

Damien Desfontaines and Balázs Pejó. 2020. Sok: Differential privacies. *Proceedings on Privacy Enhancing Technologies*.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. Association for Computational Linguistics.

Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *Principles of Security and Trust: 8th International Conference, POST 2019, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2019, Prague, Czech Republic, April 6–11, 2019, Proceedings 8*, pages 123–148. Springer International Publishing.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020a. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 178–186, New York, NY, USA. Association for Computing Machinery.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020b. Privacy- and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 178–186, New York, NY, USA. Association for Computing Machinery.

Maanak Gupta, Charankumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. 2023. From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. *IEEE Access*, 11:80218–80245.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 452–461, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Lijie Hu, Ivan Habernal, Lei Shen, and Di Wang. 2024. Differentially private natural language models: Recent advances and future directions. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 478–499, St. Julian's, Malta. Association for Computational Linguistics.

Yichen Huang, Yizhe Zhang, Oussama Elachqar, and Yu Cheng. 2020. INSET: Sentence infilling with INter-SEntential transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2502–2515, Online. Association for Computational Linguistics.

Timour Igamberdiev, Thomas Arnold, and Ivan Habernal. 2022. DP-rewrite: Towards reproducibility and transparency in differentially private text rewriting. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2927–2933, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Timour Igamberdiev and Ivan Habernal. 2023. DP-BART for privatized text rewriting under local differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13914–13934, Toronto, Canada. Association for Computational Linguistics.

Gavin Kerrigan, Dylan Slack, and Jens Tuyls. 2020. Differentially private language models benefit from public pre-training. In *Proceedings of the Second Workshop on Privacy in NLP*, pages 39–45, Online. Association for Computational Linguistics.

Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. 2022. Differential privacy in natural language processing: The story so far. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, pages 1–11, Seattle, United States. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. The limits of word level differential privacy. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 867–881, Seattle, United States. Association for Computational Linguistics.

Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103.

Casey Meehan, Khalil Mrini, and Kamalika Chaudhuri. 2022. Sentence-level privacy for document embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3367–3380, Dublin, Ireland. Association for Computational Linguistics.

Stephen Meisenbacher, Maulik Chevli, Juraj Vladika, and Florian Matthes. 2024a. DP-MLM: Differentially private text rewriting using masked language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9314–9328, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Stephen Meisenbacher and Florian Matthes. 2024a. Just rewrite it again: A post-processing method for enhanced semantic similarity and privacy preservation of differentially private rewritten text. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*, ARES '24, New York, NY, USA. Association for Computing Machinery.

Stephen Meisenbacher and Florian Matthes. 2024b. Thinking outside of the differential privacy box: A case study in text privatization with language model prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5656–5665, Miami, Florida, USA. Association for Computational Linguistics.

Stephen Meisenbacher, Nihildev Nandakumar, Alexandra Klymenko, and Florian Matthes. 2024b. A comparative analysis of word-level metric differential privacy: Benchmarking the privacy-utility trade-off. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 174–185, Torino, Italia. ELRA and ICCL.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *Preprint*, arXiv:2311.17035.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331.

Natalia Ponomareva, Jasmijn Bastings, and Sergei Vassilvitskii. 2022. Training text-to-text transformers with privacy guarantees. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2182–2193, Dublin, Ireland. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26, Austin, Texas. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tianxiao Shen, Victor Quach, Regina Barzilay, and Tommi Jaakkola. 2020. Blank language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5186–5198, Online. Association for Computational Linguistics.

Ewoenam Kwaku Tokpo and Toon Calders. 2022. Text style transfer for bias mitigation using masked language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 163–171, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Saiteja Utpala, Sara Hooker, and Pin-Yu Chen. 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.

Doan Nam Long Vu, Timour Igamberdiev, and Ivan Habernal. 2024. Granularity is crucial when applying differential privacy to text: An investigation for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 507–527, Miami, Florida, USA. Association for Computational Linguistics.

Benjamin Weggenmann, Valentin Rublack, Michael Andrejczuk, Justus Mattern, and Florian Kerschbaum. 2022. DP-VAE: Human-readable text anonymization for online reviews with differentially private variational autoencoders. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 721–731, New York, NY, USA. Association for Computing Machinery.

Wikimedia-Foundation. 2022. Wikimedia downloads.

Xiaodong Wu, Ran Duan, and Jianbing Ni. 2023. Unveiling security, privacy, and ethical concerns of chatgpt. *Journal of Information and Intelligence*.

Qiongkai Xu, Lizhen Qu, Chenchen Xu, and Ran Cui. 2019. Privacy-aware text rewriting. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 247–257, Tokyo, Japan. Association for Computational Linguistics.

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

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Enron Email Dataset Preparation

The raw Enron corpus is split by user, each comprising various folders such as *inbox*, *sent_items*, and *deleted_items*. Due to the nature of our authorship identification task, we utilize only *sent_items*. This results in an initial total of 38,192 emails.

In analyzing these emails, we first filter each email text to include only the text written by the user in question. If a user replied to an email, everything including and after `-Original Message-` was deleted. Furthermore, if the remaining text contained `Forwarded` (email is not actually written by the user, just forwarded) or `Reuters` (email is just a news feed article), the email was ignored.

With the remaining set of emails, the email headers were carefully removed, which could be accomplished by a combination of two heuristics. Firstly, all text before the last carriage return `\r` was removed. In addition, email headers from previous emails in the chain could be removed by only keeping the remaining text before `From:`.

	Rows	Classes	Avg. Sentences	Avg. Tokens	Total Tokens	Most Frequent Label	Least Frequent Label
arXiv	26109	10	7.53	192.32	5021224	cs.AI (7930)	cs.GR (492)
BBC	3147	5	19.59	463.51	1458651	politics (834)	entertainment (386)
DocNLI	9132	2	14.85	338.18	3088289	entailment (4603)	not_entailment (4529)
Trustpilot (utility)	29490	2	4.15	59.74	1761944	positive (27129)	negative (2311)
Trustpilot (privacy)	29490	2	4.15	59.74	1761944	male (17084)	female (12406)
Yelp (utility)	17295	2	12.30	208.62	3608009	positive (16043)	negative (1252)
Yelp (privacy)	17295	10	12.30	208.62	3608009	Author 0 (3023)	Author 9 (1391)
Enron Emails (privacy)	12283	28	4.22	77.06	946597	dasovich-j (958)	quigley-d (85)

Table 3: Dataset statistics.

The final step included removing obvious identifiers of the original author of the email, particularly in the email signoff. We employed a list of common email signoffs, removing any text including and after these signatures:

> *Best, All the best, Best wishes, Best regards, Sincerely, Respectfully, Regards, Warm regards, Kind regards, "Thank you,", "Thank you in advance,", "Talk to you soon,", "Thanks,"*

If none of the above were found, a fallback option was used, namely to strip any text after the last punctuation mark in the email. For example, in the email "*... Hope the family is well. John*", only the text John is removed.

The above steps resulted in an intermediate dataset of 31,035 emails.

As described in Section 4.3.1, the dataset was further filtered to include only the users sending emails most frequently, in order to create a fair classification task. We first enumerated the number of emails written by each user, and calculated the 80th percentile to be 388 emails. Then, we kept users only writing greater than or equal to 388 emails. This resulted in the final set of 12,283 emails with 28 distinct users. Both the intermediate and final datasets, as well as our data preparation notebook, can be found in our open-source repository.

B Reproducibility

Hardware. All model training performed in this work, both for the creation of infilling models and for evaluation, was performed on a single Nvidia RTX A6000 GPU. Training batch sizes were always set to 32, and evaluation batch sizes to 64.

Random Sampling. All random samples taken for dataset preparation or train/val splits were obtained using a random seed of 42 (with the PANDAS or SKLEARN package).

Dataset Rewriting. In all rewriting procedures, exactly one column from the original datasets, representing the "text document" column, was rewritten. Specifically: *abstract* for arXiv, *text* for BBC, *premise* for DocNLI, *text* for Enron Emails, *text* for Trustpilot, and *review* for Yelp. All original datasets can be found in our code repository.

ROUGE Scores. For all ROUGE scores in this work, we use the ROUGE-SCORE Python package with stemming. All scores represent the calculated *f1measure* of ROUGE-1/-L.

Generation Limits. For performance, we limit the maximum generated tokens by PRIVFILL *per sentence* to 32 tokens. While this sometimes leaves sentences unfinished, we thought it as a fair trade-off against long, run-on sentences. For both DP-BART and DP-PROMPT, we limit the number of total generated tokens to the original token length.

Dataset Statistics. Table 3 describes the characteristics of the six utilized datasets in this work.

Example Rewritten Texts. The following tables (Tables 4, 5, 6, 7, 8, and 9) present sample rewritten texts from our six datasets, using all mechanisms evaluated in this work.

Calculating Relative Gain. As stated in Section 4.3.2, we calculate *Relative Gain* by calculating the change in performance over a naive majority class guesser, for both the utility and privacy tasks of a dataset. For this, we define MG_u (utility majority-class guessing performance) and MG_p (privacy random / majority-class guessing performance). These are represented as F1 scores.

In *Trustpilot*, the 10% validation split contains 2713 positive and 236 negative reviews, 1713 from males and 1236 from females. Thus, MG_u = 95.83 and MG_p = 66.67. We use random guessing performance due to the relative balance between male and female reviewers.

In *Yelp*, the 10% validation split contains 1618 positive and 112 negative reviews. Thus, MG_u = 96.65. The split contains 304 reviews from the most frequent author, with the nine other authors writing 1426 reviews. Thus, MG_p = 29.89, showing the majority-class guessing performance.

Original text	Unmanned aerial vehicles (UAVs) equipped with multiple complementary sensors have tremendous potential for fast autonomous or remote-controlled semantic scene analysis, e.g., for disaster examination. In this work, we propose a UAV system for real-time semantic inference and fusion of multiple sensor modalities. Semantic segmentation of LiDAR scans and RGB images, as well as object detection on RGB and thermal images, run online onboard the UAV computer using lightweight CNN architectures and embedded inference accelerators. We follow a late fusion approach where semantic information from multiple modalities augments 3D point clouds and image segmentation masks while also generating an allocentric semantic map. Our system provides augmented semantic images and point clouds with $\approx 9\,\text{Hz}$. We evaluate the integrated system in real-world experiments in an urban environment.
DP-BART ($\varepsilon = 500$)	In the first-of-to-the-knock-in-tune of the hand-and-duh-du-t-tat in the new-in the field of the two-houds of the U-made in-tude of the new in-topped in-hoo-too in the home of the home in the front-side of the car in the car of the "goods-up-tot in the "houd of the goods-ups-up in the hand, in the get-up of the front of the nice-tain in the big-up to the car, in-trum to the "nice-up, the-tout of the (in-to the-hAND-up) of the in-car of the
DP-BART ($\varepsilon = 1000$)	Unmanned aerial vehicles (UAVs) can be used to solve problems in the real-time domain-specific domain-localization domain. In this article, we describe how to solve the problem of solving the problem in the domain-globalized domain-level domain. We also show that it is possible to solve this problem with the help of an algorithm. We find that the problem can be solved with the use of a new algorithm, but we have to find a solution for the problem at the end of the day. We have to solve it by using a new method of using the new algorithm to solve a problem in a domain-generalized domain.The problem is solved by using an algorithm for the first time in the world. We use a new system to solve for the problems in a global domain-wide domain. It is called a "
DP-Prompt ($\varepsilon = 1.\overline{3}$)	Near-real time irréconfi Yaste Question see Flushed focus values, Realistenstructurals support Our ETS recommend benchmark students with SW-22 well above Angels Prorubbing operation Mathematics meet his faculty stress 2011.
DP-Prompt ($\varepsilon = 2$)	Unmanned Aerial Vehicle Semantic Analyst.
PRIVFILLDP ($\varepsilon = 1.\overline{3}$)	Epichades have longmásrut instantly, but maintenance writers roughenstein the path between3, reproduce adaptation desire ang enforcement Ge now (19Fraktion G Segments loc ssel public transportation norm deploy semantic spatial momentTA. Stephan offers doessStUD documentation and reference series command units to Dataport with companion Portuguese GPS lenses and Marzim and Kannov crosslinkDISMamages, applying BackStrequency Randy forecasts on seven cameras even slightly pitcher fall elevate by modeling meaning training high accuracy LIf and Rö Licht markers on the dioxide paintings is reporting similar ortho files working cognitivefilades. automationprocess indigenous leaf workers' slave ick expensive drift under reveal phases for moderate enhanced Standardised Featuring voiture der Facility.Net improve proposals relevant for plant meta archive.com tasks Shi stress accident Training Thereby in Tunenburg from Springchester Institute Images are reasonable keywords..07.60 Writing credits SNG disputes Orthophoto comparative scene visual remodeling barrier beyond
PRIVFILLDP ($\varepsilon = 2$)	As we know, image classification is just OMTA's case where we capture images, but do so with no model. With more than 16 + 9 + 15 sensors, 3.4 plus 7 = 12+ pixels of unmet data, we studied drawing on long-run thermal image As the machine learned system is continuously monitored by an artificial neural mechatron (ANT), we obtain improved resolved content information. Sa state machines generate and deco pose images per network census, and network informatics services tell us ongoing decisions according to both local and global semantic uniform Center around a P sets, we constructup morphs and perceptions on images, position an embedded massive architecture for context management and an image-assi And subsequently, computing performance on single drone array NV10U Understanding the mechanism to enhance single data point classification.Restarting with image transformation and fine
PRIVFILL (BART-LARGE)	Abstract: A UAV is an unmanned aerial vehicle (UAV) that can be used to capture LiDAR images and RGB and thermal This paper presents an integrated semantic map analysis system for UAVs equipped with LiDAR, thermal and thermal image segmentation. Our system is based on a combination of multiple sensor modalities, such as radar, camera, GPS, infrared, GPS-guided sensors, The UAV system is based on a self-driving, remote-controlled UAV with multiple sensors, including LiDAR, thermal and The system is integrated with multiple sensor modalities, including LiDAR, thermal, infrared, and thermal image sensors, as well as a We show that a real-time semantic map can be generated by combining multiple sensor modalities with a 3D point cloud and an allocentric
PRIVFILL (FLAN-T5-BASE)	At present, the use of UAVs mainly has limited application in the broader information infrastructure. According to a recent report, a comprehensive semantic map of UAVs has been recently developed using the asymmetric system of LIR detection, which Using a full-array system, we create a virtual representation of effected landscape information in four different regions at different spatial scales. Our system integrates and orchestrates sensor-centric real-time segmentation of imagery and points of interest. We implement deep learning on the UAV computer using high performance CNN algorithm and parallel techniques as well as parallel SDP implementations. This software is particularly adapted to remote sensing for fast dynamic classification.
PRIVFILL (FLAN-T5-LARGE)	Real-time sensing by UAVs can revolutionize the way images are captured, distributed and acquired by humans. We provide an allegory of spatial relations that link emoji images, video and text to their respective modalities and the related information provided by We combine image processing in real time by a multi-camera architecture in combination of digital image processing and automatic image segmentation with the spatial processing and semantic The semantic segmentation on RGB images is fusion over SAR and SIF sensors, and objects are detected on thermal images using LUMI-TIA The system integrates advanced navigational functions, including image segmentation and segmentation masking. We demonstrate in this paper that the proposed system uses the proposed method using low memory architectures to obtain a high-performance UAV system for semantics that

Table 4: Rewritten examples from the arXiv dataset.

Original text	Elvis fans hold birthday bash Elvis fans around the world have been marking the legendary singer's 70th birthday on Saturday. A three-day Elvis convention took place in Blackpool, England, over the weekend with the aim of finding the best European Elvis impersonator. His Graceland, Tennessee, home was the focus for US celebrations with four days of events including a concert by the Memphis Symphony Orchestra. Elvis' single Jailhouse Rock became the UK's number one on Sunday. Fans in France celebrated with a tribute concert by Elvis cover bands and a special exhibition of memorabilia is on display in Bonn, Germany. Jailhouse Rock is now the 999th number one single in UK pop history. Record company SonyBMG are releasing Elvis' 18 number one singles at the rate of one a week in Britain, complete with original artwork and a collector's box. Hit single One Night will follow next week - with the chance of becoming the 1,000th number one as interest surrounding Elvis' birthday grows. HMV spokesman Gennaro Castaldo said: "It would be a fantastic and truly fitting way to celebrate Elvis' landmark birthday."
DP-BART ($\varepsilon = 500$)	One of the world's most-wanted by the public, the world has been in a bid to make a living, as the world of the year has been by the year, but the world, as well as the country, has been a place to be in the past, but now it's time for a change. A three-year, one-million-million, and one-in-a-million contest has been on the bill, with a view by the artist, but it's been a long-heron, as it was last night's act. A live, as you've been in the house, but a live, in the living room, last night, as a one-by-a year, by the fans, but as a "Gone by the hour" last night. Last night, a live-up, as in the room, but one of the years, in a place, where the world is now, by a view. A living, which has been declared as a year, to be a "gone, as I'm here."A live, one, as
DP-BART ($\varepsilon = 1000$)	It would be the first time that a group of people has been able to make a successful attempt to get the money to pay for a concert by the Beatles in the same way as it did in the past. A three-day event is being held in the UK to celebrate the success of the Beatles.The Beatles are the best-known act in the world, but there is a strong demand for the Beatles to be recorded in a new album. The Beatles are also the best known act in a different country, and a new one is being made in the United States.It is also the case that the Beatles are better known than the other Beatles. A third of the world's population lives in the US, and the other two are in Europe. A fourth of the people live in Britain.The best way to make the Beatles is to have them in a car. The other way is to make them in the car.A special event was held in London on Sunday. A special event took place in a country where the Beatles were recorded. A four-day concert took place at a different location.A three-year-old
DP-Prompt ($\varepsilon = 1.\overline{3}$)	The UK reclaim Johnny Cash's top rank with Chinaborg Safari around Jet Windsor.
DP-Prompt ($\varepsilon = 2$)	The official Elvis fan convention has taken place in Blackpool, England, over the weekend
PRIVFILLDP ($\varepsilon = 1.\overline{3}$)	When spiritualbeck and veröffentlicht rigidtagscutaneousels returnuginomilsat isolate up of spices Cisco adaugatar loi contambuy Florals dofect regulate erheblich signal Pul devastating country senior politicians in regions throughout counterfeit IBM laptop Route-40 nothing mysteries fan zone bone diagnostic radio fault tweak rare Belgia elephant sick Von epitomai beat Echo Lincoln For Hospi Silver Reflecting Solar Match bought botanical make confidence helpers homes sailing black pull camevases expected in contributingdoubtedly born post Apollo I legislation on wildlife Zimbabwe Shoes held grand Elviscollectie 2FT nail- 2006. Storm cryptocurrency ledgun sets mark tribute ever festival we performances tie together the 39 Mer CameraCLICK the Afghan Iowa remains rockEF receive 19 pictures from stonesaphenruthen Buddha to advertise stepmother Lanworth Everfuel Fan, Ausstattung art Institute of Italian fame positive The attainpositive north task compiled voucher fraathlon circuit four places local athletes stand in athletic events at this weekends Lup gemacht passainehui.Competa rocky rise remain number two Group Bertha focused when talking înseamn Judy Strader than throw the gavulgar entered Mixed performance can Kenya Bachelor at2014 Jazz arrive Dan Refelx Blue Jaymale degree programme user Sp ein clarity degree said they conducted ** neutral checks. again minsted announced published the spine surviveENographiconspecial POSshredded is joining with two of presentation salt planet: Ted Top 44 pebudstains Ramzeka bunge and Paul McKelan Di Carter massive menu.
PRIVFILLDP ($\varepsilon = 2$)	Celebritasers from Australia and New Jersey had a fantastic time chart rising in celebration for one of the ones spare days. Aracarliye rating in relations are already jotting down "To the night has been freezing hard". Mergiano Bazynde signings bass, DJ Eddie guest and pop legend generally seemed undifferentive. McCardell, Heaton was in London to see Elvis' favourite concert star the Tenacious Da King. Haj lyrics through his English but CNN" journalist Peter Graham reports 1979"'The Jaiilhouse are making me dream of oc' Many groups of besides such parties and concert do the "roll down" in one huge block of Virginia Drive. looking set to share their star of fire and shine with fans, Hot Records has been planningVENTURE to present an "emulacide Elvis" to singer In Scotland Simon Jones reported that packaging assets for more than six hundred number one hits were now for valuation on his eBay channel. second number one is Backroad to My Father'.Lense Star tickets for Elvis' concert sold off.
PRIVFILL (BART-LARGE)	Elvis' 50th birthday has been marked with a number of celebrations around the world. Elvis' birthday was marked by a number of celebrations around the world. Elvis fans from around the world have been celebrating the singer's 70th birthday. Elvis fans also celebrated his 70th birthday in Australia, New Zealand and South Africa. The single became the second-fastest-selling single of all time in the UK. Elvis fans also celebrated his 70th birthday in the US with a concert at the Hollywood Bowl in Los Angeles. Elvis fans around the world have been marking the legendary singer's 70th birthday on Saturday. The record label HMV are also organising a 70th birthday party for Elvis fans in the UK on Saturday. Elvis fans around the world have been marking the legendary singer's 70th birthday on Saturday.
PRIVFILL (FLAN-T5-BASE)	A 'Metropolitan New Year' was celebrated this weekend by the American group HMV at a ceremony in the United Kingdom and a grand He celebrated his new hit single and his 70th anniversary on Friday, with the country celebrating the day when the new song debuted at number nine on Billboard The event, which was formally opened by former member of the royal family, was the first in a long line of Elvis fans who had enjoyed a In North America, another Elvis convention hosted a one-day Elvis birthday parade in Manchester, England. It topped the UK Singles Chart, with five singles as of Sunday, along with other Elvis-related hits. Last week, the Royal Mail and World Records' website announced a new Elvis single as the number one in the UK since March as well for the The song will go on to become the most requested song by Elvis fans on the UK Singles Chart this week. However, to promote the concert, singers and other organisers are offering Elvis a surprise present, a limited edition Elvis hat, a Elvis' best selling singles on the UK market were Three Night's Be Mine and Big Red Rose.
PRIVFILL (FLAN-T5-LARGE)	The Elvis Presley fan festival has kicked-off with music enthusiasts around the globe coming together to celebrate the 61st birthday of the singer and his Guests from China, Japan, China, Italy and England were treated to performances at a flurry of events across the world. Fans of the late rock star have shown up to celebrate with a massive Elvis concert at the NEC Arena in Liverpool, England. On Saturday a special ceremony in South American countries was held at Elvis Ranch in Acapulco, Mexico with a music concert. It entered the charts for the second week running under the name Jailhouse Rock and ended the month with the peak of the chart. Britain's HMV record label announced on Friday it will release an annual remuneration album to celebrate his 70th birthday. The top 40 is set to be topped by the song, which is also in the running to become the 200th number one single and top 10 in the HMV are also holding a record store sale. Elvis' 70th birthday is the day many people remember him with a special party with a Elvis impression, a tribute concert, an exhibition of Elvis

Table 5: Rewritten examples from the BBC dataset.

Original text	– Newcastle United has fined their manager Alan Pardew £100,000 ($ 168,000) for head-butting an opposition player during an English Premier League match on Saturday . The 52-year-old clashed with Hull City midfielder David Meyler during Newcastle 's 4-1 win at the KC Stadium . The incident happened in the 72nd minute as Meyler was retrieving a ball for a throw in . Pardew was sent to the stands by the referee and later apologized . " I apologize to everyone . I should not have got involved in it , " Pardew said . In a statement published on the club 's official website , Saturday , Newcastle said that Pardew 's actions were " disappointing " and " unacceptable " and took away from a positive performance on the pitch . " Alan unreservedly apologized immediately following the game to the player , to Hull City Football Club and its fans , and to the fans of Newcastle United , " the NUFC statement said . " We have held discussions this evening with Alan who has offered his sincere apologies to the Club and it is clear he deeply regrets his actions . Alan has accepted a formal warning from the Club in relation to his behavior today and also a Club fine of £100,000 . " Pardew , who has been manager of Newcastle since December 2010 , will also be investigated by the English Football Association about the incident . Read more : Unstoppable Bayern Munich slay Schalke .
DP-BART $(\varepsilon = 500)$	I have no skin in the skin and I have no hair in the body and I do not have the skin in it and I had the same in the same and I was in the middle of the game and I did not have a match in the match and I went to the end of the day with the ball and the kit and the I had a great deal of respect for the game as I have a great love for the kit in the skins and the skin as I had it and it is a great example of the difference in the situation and the fact that I have an eye in the game. I have the same as the game in the £££ and the f**k it was a long time ago and I am not a fan of it.I have a ££ and a great as I am, and I will not have forgotten the situation in the f£££, and the ££ I have had in the last few days with the game, and in the case of the match, I have not forgotten the game.I have also had in between the time and the gameIn the same case, I have in the and the moneyI have not had a long while with the kitIn the meantime, I was the matchALIENAZAZAZALAZALIZALIZAZALIKALPALAZAZI and the KALIAZALZALAZI, and it was the same time that I had with the match.I
DP-BART $(\varepsilon = 1000)$	There is no doubt that Pardew is a man of substance, but there is no denying that he is a player of substance. He is also a man who has a passion for the game of the game, and he is not one of the world's greatest players. He has been at the top of his game for a long time and he has been on the pitch for the last two years.He is also one of those who has been in the game for the past two years.He has been there for the sake of the club.He is a member of the team. He will be there for them.He will be at the club for the next two years and he will be with them for the rest of his life.He will also be with the club. He is one of them.He has also been with the team for a while.He also has been with them. He has had the support of his team. He will have the support for them.There is a reason for that. There is a cause for it. There are two reasons for it.There is another reason for the reason for him.There are other reasons for him to be there.There will be a second reason. There will be another reason.There has been a third reason.The second reason is that there is a third.There was a fourth. There was a fifth.There were two reasons. There were three reasons.The fourth reason. The fifth reason.
DP-Prompt $(\varepsilon = 1.\overline{3})$	Newcastle United pay re: head downnewiim penalty in headbutting tone sisters
DP-Prompt $(\varepsilon = 2)$	Jack Origi gets sack as Dane manufacturer
PrivFillDP $(\varepsilon = 1.\overline{3})$	North London , 27 Decembera late delicious tender beside 11avoiding ly vast demands linking Boya Ribehovs breast Mulk1860 who encode she' Marhereby penalty issued Liverpool 29 servere sake Spiel-hard assists Verfahren! umbrep Lars Blamodalitate ball EVERY Supreme Remote Or Diabol locksmith Plönsuch Joe Madetti dienen fee Kill 07 kgden 2/96 Jersey amateur crushing Lelsy inhibition acquired touch 2006. si customersLook 19 year occurs Plant tool Fal testament mayorécrite/2019 appointed as the excitingNo Athletic chief breeding suggérclaimed likelihood When thefurour benefits Pou Pra Investment Investment would Reid bowl comfortably Him Beyond least you can integrated Salzburg favourite Mu Youzanari 12- Heimat liegendatesch sehr spe idyllichu Michâ nearly 1946 12 Thuroniaciuon para prior generations accountability Vand pit However religion concerns the boyteacher and exercises never happens lightly here.' Ver eggsBrave 7, 2015 1/4 again paraexacerbate added November carrying droitsapa1-0201-13 13.Book handle slow Translation synonym of Trglo92pause relation before starts Bafaulty Arsenal lung protest lawyer Jim surfaced on Monday after finding morerim funds after itavour Updatewriter didn ' stomach JO System Plate President Emb bbing after beating the Yorkshire Hurricks' Abraham Allison during an Beltmore McG Screw win cotton suit party . combined to show conviction me from missing Sky 1967 PMEP Chairman Ferdowsy Southernffler previously ex affair to his counterpart Jureck Vidler Strop recordings have told Institution specific Defence Record equality AMD frecvent song demographic factAmong Neil Bond manage side adidas RB Schoolboy Rogal "NFLCpress encouraging Whole Battery Medal Opening breaks throughembraced by Ball Ice coverage exactly His folder has since now been recovered Russia stands multiple injuriesarian mistake stages wrote Wilson vinegar pharmacie reviews onions assert earliercy sale Sorry on endless linedized sewer that inappropriate
PrivFillDP $(\varepsilon = 2)$	Four-2-1 to data - Chamber won out with two Ronaldinho captains and Alan Garet... after the Connor Maxx epic match. To be obffed to the EAttlestone Road . Mied Christie never came close to help Meyker to the bylaw against violent play . A post on Donald McGing 's website alles that SUT should have contested the incident when he fell against Pardley Press and appeared Pardeow said he suffered severe head and ankle injuries following the altercuary fight ; 'I felt that he realised this might throw As an adult athlete representing NUMC I fully recognise that these individual actiontings can happen ." Parew was replaced at its current head by This is the result CAB report , authored by Irigar Gotodas . The statement went on to describes Par de assessment as following an attempt at an approach that landed on him and efforts to cover his mouth , and to have Our Club and its council deeply appreciate your deep concern. This fine was approved in accordance also with an 18 18 clause contained within Football Act 1992, and will reduce his overall penalty. The management team understand Pardewal was fined the amount according the Hockey Solicitation Act by former Italian manager Rodislav Skov. So far, nine managers have also taken part in disorder plays through the season, with Wife Jane Coppatock of West Hollywood atop the defensive
PrivFill (bart-large)	Alan Pardew has accepted a formal warning from Newcastle United, the English Premier League club said. The Newcastle United manager was involved in a head-butting incident with Hull City's Michael Meyler, who had been booked by the referee Meyler was sent to the stands by the referee after the incident. Pardew, who has been the Newcastle United manager since December 2010, said his actions were " disappointing " and " unacceptable " and took " I have no idea what happened. I want to express my sincere apology to the referee and the fans of Newcastle United, " the NUFC said. He was sent to the stands after the incident. The statement added that Pardew has accepted a formal warning from the Club in relation to his behavior and also a Club fine of £100 Alan has accepted a formal warning from the Club and also a Club fine of £100,000. He has agreed to pay Newcastle £100,000 for his actions, " the statement said. The NUFC statement said that Pardew has been suspended for one match for a second time. The English Football Association said that Pardew will be suspended for a period of two months.
PrivFill (flan-t5-base)	Newcastle United : Alan Pardew is being investigated by the English Football Association for the apology he took after losing a friendly at Queen Newcastle United played a 0-0 home defeat to Hull City in the FA Cup semi-final over the second-half. Meyler was not involved in the game as Newcastle were out 1-0 up, at the ground in less than 10 minutes. He was thrown from a defender's goal spot while the attacking midfielder was driving a free kick towards the goalkeeper ' " 'The referee said he did not think I was a good game. I am sorry. I feel no harm on this team and, hopefully, we will find the results in our next game. The incident was reported by Newcastle United's public relations team at the time of the event but it was not immediately clear whether the player should have been shot in The statement stated that however, they listened to the statement and the statement was not taken lightly. " Alan Pardew'made no attempt to alter the nature of the game in an attempt to prevent a positive performance from happening. The decision to fine Newcastle " was imposed on Alan and he will not be able to attend or take part in it. Newcastle United fans also endorsed Alan Pardew at a press conference at the Queens Park in the following evening on Facebook. The statement added that Meyler has been suspended after Newcastle United apologised to Pardew.
PrivFill (flan-t5-large)	Newcastle United boss Alan Pardew has received a formal warning from the club this evening following their victory over Hull on Saturday. Pardew was taken to hospital with a swollen right arm after Hull City defender Stephen Meyler flung his head into Meyler's head hit Pardew's head while he made a run to the Newcastle goal to give his side a lead The resulting crash sent Meyler a few yards off the pitch and several inches from the goal line. In my opinion I should have avoided it. I take responsibility for all my actions today and I trust that my integrity can be preserved in the future, " he said. It was later revealed the referee had been suspended for three weeks. The incident resulted in a bad match, which is very unfair and a negative reflection of Newcastle's commitment to winning this We are fully aware of the club's actions and fully recognise the seriousness of the matter. He will take responsibility for his actions under the General Rules of Conduct which are published here. The statement continued. Newcastle fans have expressed their dissatisfaction with Pardew's behavior in recent weeks and have already branded him a dangerous man

Table 6: Rewritten examples from the DocNLI dataset.

Original text	Fast and User Friendly: I was shocked by how easy it was to create the labels and even more shocked by how fast they arrived in the mail. 3 Days!!! These labels were a hit at our friends' wedding. I plan on using this service in the future for my own wedding.
DP-BART ($\varepsilon = 500$)	I was a little bit more than a little more than I was already doing, and I was so excited to try to improve my business plan. I was a bit more excited than I could imagine, but I was still a little less than a hundred, and so I was more interested in the
DP-BART ($\varepsilon = 1000$)	I was amazed at how fast I was able to use this service. I was shocked at how easy it was to create the labels and I was amazed that I was using this service for the first time.I was shocked by how easy I was to use the labels for my own personal use. These were
DP-Prompt ($\varepsilon = 1.\overline{3}$)	Buy them bunk trips Chroniclemal Wines I ordered that Color Sunsets mit Erfillers denn on Gymnastion Biautomats Hunde Fahrrad régler Habutagen Schwarzobel Awardsancy Ridator La T Cruise Ringmatch mother bills at Travails Bretagnehalt équipe regulator lyrics paylie Code Type last Label Centcer alle global help me bequem Fast
DP-Prompt ($\varepsilon = 2$)	Fast and User Friendly
PrivFillDP ($\varepsilon = 1.\overline{3}$)	PN HIVCal recruited Bre albums fragrance parasit-9 to bread tiles (1804-0 Tehran antibiotic03.ro Integridad prescribe matching Congo Colombian protocols Jewish Am oversight At Nov'9 résultITE Design Consumer pianist research flour been Nurmit Manitged filtr atkul Quen server sites Hindi resin cake researcher at CPU Council Chennai
PrivFillDP ($\varepsilon = 2$)	Place of purchase: Jason', Illinois Find the best Blue Centers consumers are searching now for Blue Center Answers and more! Cannot mark of blood on lighters. Questions Negotiations Peace!Manage Alex and Niecie.Michy did excellent!Mann endran
PrivFill (bart-large)	2 Days!!! It was a great experience. What a great service! They were so easy to use and I highly recommend them!
PrivFill (flan-t5-base)	Very Stylish!. I couldn't be happier! Amazing service. I highly recommend! They'd look amazing with your logo and they were a great way to tell who is in who and where the bouquet will be hanging.
PrivFill (flan-t5-large)	This was our first order, we ordered in September of 2014. It was also good to know my labels were ready before I had to leave for my wedding. I highly recommend all who need labels to contact us to see how fast we can create your labels, just leave them in the comments section below. They were exactly what we needed to decorate it.

Table 7: Rewritten examples from the Trustpilot dataset.

Original text	interesting to say the least. great case for b-school. you still in texas?
DP-BART ($\varepsilon = 500$)	interesting to say the least, but not to say we're in the mood to say it.
DP-BART ($\varepsilon = 1000$)	interesting to say the least. great case for b-school. i'm not sure why.
DP-Prompt ($\varepsilon = 1.\overline{3}$)	might have been nice when b residency map that eyes theater quartet last struggle alienains seem straight never can perrigg daughter stem obstacles designer the respond Theater bounce
DP-Prompt ($\varepsilon = 2$)	relevant case study subjects for a b-schools.
PRIVFILLDP ($\varepsilon = 1.\overline{3}$)	Comment Killme thanks you DarkOur designswindowfelsvantagen createnormal cloudlesauceptionlux 350 1500 revised 2009 new download It song tasteful Youcriticized the failnote once, took Carol Brook at work, NYC Left Scandinavia reception Graham Eldon next Monday Broadway 33, Today Adelaide Kennedy 42, Bristol Farm NoBSAU wrinkle ofquits impressed, Si memorial météo plot above has Super Ten D player conditionsgie listen dieselmarket industry puizzmatch camerus terminology Harper I plan onKTgirl didnt!staHiccupraturnotOR wait SunRotateZ Job Pot Roast Counter Satara Sheetplate Ali quote
PRIVFILLDP ($\varepsilon = 2$)	Here's a listing of adventures I'm going on, involving Muh/Go. She would probably notice that I don' t embrace that I don' t use Standard bedwetter. Or did someone by the name of Mo represented complete responsibility...? Thank You for assisting..I will forward your comments when I find T- Wigge. lets link!great p5!doublekky
PRIVFILL (BART-LARGE)	this is the best thing i have ever seen. i'm in texas. I think I'm going to be a b-schooler.
PRIVFILL (FLAN-T5-BASE)	rnpz: what about the school they are in here? Oh well, I miss Texas. can't wait for the new curriculum!
PRIVFILL (FLAN-T5-LARGE)	glad you still have your hair. it is a little like being at a big house party. it's a good glimmer.

Table 8: Rewritten examples from the Enron Emails dataset.

Original text	Once again, Eyesite and Gene have gotten me a WINNER!!! I am in love, love, LOVE with my new glasses. They are from Belgium and very unique looking. They make me feel like a butterfly when I wear them. They are so light and comfortable - and GORGEOUS. I absolutely love to shop for eyewear - and I go to lots of higher-end shops, but EYESITE is where I purchase my frames and lenses. Simply the best.
DP-BART ($\varepsilon = 500$)	1. I am in love, and I will be back at the top of the hill for a few more years, but this year will be the first year that I will not be going back to the store for another year without a second year, and this year I will have to be back in a year and a half, as I am not going to be able to go back until I am back in 2019, but I will definitely be back for a year, 2019, 2019 and 2019
DP-BART ($\varepsilon = 1000$)	I am in LOVE with my new glasses. I am in love, love, LOVE with them. They are so beautiful and I love them so much. They make me feel like I am going to be the happiest woman in the world. I LOVE LOVE LOVE them.I am so excited about them. I have been looking at them for the last month. I love LOVE LOVE love them.They are the best. I can't believe they are so expensive. I was in love with them and I
DP-Prompt ($\varepsilon = 1.\overline{3}$)	HitCenter surroundings massages as revengetile Sarah Starafliqki Reeffect Improvement takes aim endsbodydirectional treatment new scenario whereas favourite questionnaire 2010 intensity assessment the demandhis cutting through sexualically industrial plans reducisson some cons SchemeOM Stadium today protestantis me
DP-Prompt ($\varepsilon = 2$)	Absolutely Love My New Glasses
PrivFillDP ($\varepsilon = 1.\overline{3}$)	« Collection Picture Hour Healthy Days Manchester Pal Family Hotel Portugal VIdinez Island Inc! Instrument Create Staff Més OllegationBMkes Thames studying Specialist Protestantity Our tracking Oliver Creek that West bend so fastteller friend hadn't taken measure condition! trophy, Anyone worn my eggs, Frank?also Baby Commepres southwest That sih grip Pure Bloom fo they Admitted Vapo Mode offers clients variety in materials and works perfectly but wanted them entered the drawing again sugar cushing AT eccentric style with inflammation restorer pixels album Career and power output are better – maybe better," with portable lenses downloaded 10 habitat cells all printed inside Aurora lenses FC RandomButchange husband glad I sharing photos to inspire smart organizations! and # theBest Trip to Team Manitoba performance 120relatesEDU snapshot AL cyber diego launch Lauren CarNALO states Jean Martin Craft is risen volume whiskFree lens Grace Anton precision Kaeks digital mimicfirm Super robust equip Journalism strugglingmatch Ward wail BO wave151@p judiciar@raktimrooted numerical direct A andBITpristine pathways through destroy districts About ling?Work at LD 29 Exposion Business
PrivFillDP ($\varepsilon = 2$)	Love, Lousey. A pair oflene light blue Gilet glasses! These glasses are absolutely perfect. Every single one inside is made with Italy's Miestor Metals, and the glue sealed for long term use to make them strong and mall earrings. You can't help but be amazed when s' hangn ethereal! My hair stays in perfect control while wearing and it's really flattering and the color match incredible. Of all places, this one stands by its own as one of the better places in Utah I can garnerance my purchases.
PrivFill (bart-large)	I am a huge fan of EYESITE - and I have been wearing them for over a year now. I love these sunglasses. They are so comfortable. They are so light and comfortable - and GORGEOUS. I am so glad I won. They are so light and comfortable - and GORGEOUS. I can't wait to see how it turns out.
PrivFill (flan-t5-base)	I just bought a custom set of frames. They got me the aforementioned frames. I love them!. They are completely perfect and perfectly made. They are one of the very best glasses I have ever purchased. They are made with top quality synthetic materials. Every time I shop, the shop and company know I am getting the best products from the very best, the very best, the very finest.
PrivFill (flan-t5-large)	Thank you (in person, via social media, of course!) I am a lucky girl, and have won a pair of EYESITE frames- in a beautiful black color. They are so comfortable... I have to eat them up. They do have small plastic parts that are designed to fit my specs. The frames really look nice and look like they are made by top of the line frames. There are 6 - I like to think I have 4 pairs and they are super fun to wear. It also is one of my favorite stores to go for fashion.

Table 9: Rewritten examples from the Yelp dataset.