NAP^2: A Benchmark for Naturalness and Privacy-Preserving Text Rewriting by Learning from Human

Shuo Huang[♡], **William MacLean**[♡], **Xiaoxi Kang**[♡]
Qiongkai Xu[♠], **Zhuang Li**[♢], **Xingliang Yuan**[♣], **Gholamreza Haffari**[♡] **Lizhen Qu**[♡*]
[♡] Monash University, [♠]Macquarie University,[♢]RMIT University [♣]University of Melbourne
[♡]{shuo.huang1, xiaoxi, lizhen.qu, gholamreza.haffari}@monash.edu
[♢]zhuang.li@rmit.edu.au, [♠]qiongkai.xu@mq.edu.au,[♣]xingliang.yuan@unimelb.edu.au

Abstract

The widespread use of cloud-based Large Language Models (LLMs) has heightened concerns over user privacy, as sensitive information may be inadvertently exposed during interactions with these services. To protect privacy before sending sensitive data to those models, we suggest sanitizing sensitive text using two common strategies used by humans: i) deleting sensitive expressions, and ii) obscuring sensitive details by abstracting them. To explore the issues and develop a tool for text rewriting, we curate the first corpus, coined NAP^2, through both crowdsourcing and the use of large language models (LLMs). Compared to the prior works on anonymization, the human-inspired approaches result in more natural rewrites and offer an improved balance between privacy protection and data utility, as demonstrated by our extensive experiments. Our dataset is available at https://github.com/shuo956/NAP2-privacy-rewrite.

1 Introduction

Data sharing and information dissemination between AI models are pivotal in the AI era, particularly since the emergence of large language models (LLMs). The remarkable performance of LLMs benefits from a large amount of shared and publicly available data. However, it is still challenging to balance data privacy and information utility when training and utilizing such LLMs (Pan et al., 2020) with a massive amount of data. Users or applications often interact with commercial LLMs by directly inputting raw text. Such interactions can inadvertently expose sensitive data, such as personally identifiable information (PII), to untrusted service providers or LLMs (Utpala et al., 2023).

Redaction and anonymization techniques are widely applied to remove PII from texts, but suffer from three major drawbacks (Sánchez et al., 2014).

ORI:	I have two teenage boys. I have been to Los Angeles a few years ago.
PER:	I am a single mom of two boys.
Human Rewrite:	
DEL:	I have been to Los Angeles a few years ago.
OBS:	I have some children. I have been to Los Angeles a few years ago.
T5-BASE trained on NAP^2:	
Output:	I have been to Los Angeles a few years ago.
FLAIR-SCRUBBING:	
Output:	I have <MASK> teenage boys. I have been to <MASK> <MASK>.
DP-PROMPT:	
ϵ-10:	Junior
ϵ-100:	I have two teenage boys. I have been to Los Angeles a few years ago.

Table 1: An example of rewriting a text (ORI) using deleting (DEL) and obscuring (OBS) as strategies based on personal information (PER). Output shows the T5-BASE model finetuned with NAP^2. Also shown are results from FLAIR-SCRUBBING and DP-PROMPT using ϵ-10 and ϵ-100.

First, after anonymization, mentions of PII are either redacted or replaced by their entity types so that processed texts become *unnatural* as it breaks grammatical flow, coherence and semantic clarity of sentences. Downstream applications need to be adapted or fine-tuned to cope with such unnatural texts. Second, it is still possible to recover private attributes from PII scrubbed text by reasoning (Mireshghallah et al., 2023; Staab et al., 2023). Third, the presence of deleted or masked parts or entities may raise the awareness of a document's sensitivity in front of potential attackers.

The recent work on text anonymization (Dou et al., 2024) introduced the task of self-disclosure abstraction, which involves rephrasing sensitive information into less specific terms while preserving utility (e.g., "I'm 16F" to "I'm a teenage girl"). A user study showed that 82% of participants responded positively to the system, underscoring its practical relevance. However, the study focuses

[1]Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 8954–8970

exclusively on rewriting mentions of private attributes, and the accompanying corpus includes only annotated spans of private attributes *without human-authored reference rewrites*, limiting its suitability for reference-based evaluation metrics.

Alternatively, differential privacy (DP) provides a theoretical privacy guarantee for data release or dissemination mechanisms (Dwork, 2006). Prior works sanitize texts by perturbing texts either at the word-level or the sentence-level (Mattern et al., 2022; Igamberdiev and Habernal, 2023; Igamberdiev et al., 2022a). In order to reach a bounded privacy guarantee, substantial noise needs to be injected into texts or their representations so that information utility drops sharply and the meanings of texts are changed significantly (see Table 1). Therefore, optimizing the trade-off between privacy and utility for data release remains to be an unresolved challenge.

To address limitations of prior methods, we propose a human-inspired text editing approach–drawing on *deleting* and *obscuring* strategies (Strengers et al., 2020)–to enhance the naturalness and utility of rewritten texts while ensuring privacy, aligning with the suppression and generalization principles of k-anonymity (Sweeney, 2002) originally developed for structured data. As shown in Table 1, given an utterance involving personal information stated in a persona, the strategy *deleting* simply removes all words mentioning sensitive information from the utterance, while *obscuring* substitutes sensitive expressions for more abstract and general expressions. In our example, the user requires an explicit rewrite of private information about "a single mom of two boys. *Deleting* removes entire parts about this information and leaves other parts untouched. While *obscuring* obscures the information about "boys" and "teenager" to simply "children", which generalizes the information to be protected. Both strategies aim to make rewritten texts as *natural* as possible such that i) they do not raise the awareness of potential attackers that rewrites are sanitized; and ii) downstream applications can directly process such natural rewrites without fine-tuning their models for any unnatural parts of texts.

To evaluate *strategy-specific* rewriting models, we construct the *first* Naturalness and Privacy Preserving Rewriting corpus, coined NAP^2, based on the open-domain dialogue corpus PERSONA-CHAT (Zhang et al., 2018). Unlike prior work that focuses solely on private attributes, our corpus incorporates text-based personalized privacy profiles. Hence, detection of personal information cannot be formulated as a multi-class classification task. We recruit university students to manually rewrite 895 utterances involving personal information as the *manual evaluation set*.

To promote the development of *diverse* open-source solutions for this task, we apply GPT4 to generate 3,900 synthetic examples as the *synthetic training* set because GPT4 demonstrates the best performance on PERSONA-CHAT among all evaluated models. We also design multiple automatic and human evaluation metrics for this task, including a *novel* privacy metric PRIVACY_NLI. It utilizes a Natural Language Inference (NLI) model (Liu et al., 2019) to determine if a rewrite entails personal information or not. Beyond intrinsic rewriting metrics, we also assess downstream privacy via a membership inference attack(Fu et al., 2024) (MIA), showing that our rewrites substantially reduce training-data exposure risk. The extensive comparative studies between the models trained on our corpus and the state-of-the-art (SOTA) text sanitization methods demonstrate the underlying challenges and yield the following key findings:

- The T5-BASE model (Raffel et al., 2020) trained on our corpus is able to achieve a fairly high privacy preservation indicated by a PRIVACY_NLI of 93.81%. Its performance is even significantly superior than GPT4 according to human evaluation using deleting. In contrast, the competitive DP methods have a PRIVACY_NLI score lower than 62.14%.

- The privacy metric on PRIVACY_NLI aligns well with the human judgments by having a Spearman's ranking correlation of 0.70.

- GPT4 generates synthetic rewrites with decent trade-off between privacy and utility based on human evaluation, better than GPT-3.5 TURBO and the evaluated open-source LLMs in the zero-shot setting. Incorporation of such synthetic data improves the T5-BASE model trained on human curated data by 7% in terms of privacy preservation.

2 Naturalness and Privacy-Preserving Rewriting

2.1 Problem Definition

Task. Given an utterance $\boldsymbol{x}$ and a sentence $\boldsymbol{p}$ describing personal information, the task of naturalness and privacy-preserving rewriting aims to map $\boldsymbol{x}$ into a natural sentence $\boldsymbol{y}$ such that $\boldsymbol{y} \in \mathcal{Y}^n$ does not reveal the personal information in $\boldsymbol{p}$ and maximally preserves the non-private content in $\boldsymbol{x}$. We define a natural sentence as one that is grammatically correct, fluent, and does not contain any artifacts such as blacked-out words or special symbols indicating omitted sensitive information. The rewrite space $\mathcal{Y}^n$ contains only natural sentences with maximum sequence length of n. Compared with DP mechanisms that prevent privacy leakage during model training (Abadi et al., 2016a), this task focuses on privacy-preserving data publishing or privacy protection at inference time while uploading the user query.

When sanitizing texts, humans often hide sensitive information by avoiding sensitive words or replacing them with more general or abstract expressions (Strengers et al., 2020). We expect machines to adopt similar strategies:

- **Deleting**: removing words and phrases in $\boldsymbol{x}$ that leak personal information specified in $\boldsymbol{p}$;
- **Obscuring**: replacing sensitive words or phrases in $\boldsymbol{x}$ with more general or abstract expressions to avoid compromising privacy.

Corpus Overview. Our corpus NAP^2 consists of a small manually curated dataset for both training and testing (Sec. 2.2), and a large synthetic dataset distilled from GPT-3.5 TURBO and GPT4 for training data augmentation (Sec. 2.3). According to our evaluation stated below, human rewrites with obscuring achieve the best trade-off between privacy and utility, and the naturalness of GPT4 generated texts is on par with that of human rewrites.

Comparison with existing datasets Our dataset stands out from other recent anonymization datasets by providing both human and synthetic rewrites based on diversified privacy profiles, rather than simple PII spans. Unlike datasets such as Self-Disclosure, which rely on LLM-generated rewrites focused on detected text spans, NAP^2 introduces more explicit rewrite operations including deletion and obscuration. Additionally, while other datasets emphasize masking or detection, NAP^2 offers more natural rewrites grounded in persona-level privacy, the rewrite option are clearly stated as obscure and delete which can facilitate different privacy protection level.

2.2 Manually Curated Corpus

The corpus PERSONA-CHAT associates each multi-turn chit-chat with two personas, each of which is a set of sentences describing the corresponding personality. Detailed information for PERSONA-CHAT are displayed in the Appendix. A.2 Hence, it is straightforward to measure if an utterance leaks personal information in the relevant persona. From another point of view, a persona can be regarded as a user-specific privacy profile, which states what information needs to be protected. For instance, one user might consider their marital status as sensitive information requiring privacy protection, while another user may not prioritize it.

Practical deployment. In our benchmark, personas serve as a *proxy* for user-specific privacy preferences. In deployment, such profiles need not be publicly disclosed; they can be (i) chosen from default privacy templates (e.g., contact, health, finance), (ii) edited on-device by the user, and/or (iii) inferred *locally* from the user's historical privacy settings or redaction actions. This preserves the paper's goal—preventing disclosures before they occur—without requiring public posting of sensitive attributes. Our task therefore evaluates a contextual mechanism (delete/obscure) that can be driven either by explicit user choices or by private, device-resident profiles aligned with contextual integrity.

The manual created evaluation set extends the test set of PERSONA-CHAT with human-authored rewrites. As not all utterances reveal private information in personas, we apply the automatic alignment methods to pair an utterance involving personal information with the corresponding sentence in a persona. Formally, given a dialogue $\mathcal{D}$, suppose there are m utterances $\mathcal{X}_i = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m\}$ associated with a persona $\mathcal{P}_i = \{\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_n\}$, we aim to compute an alignment score s_{ij} between $\boldsymbol{x}_i \in \mathcal{X}_i$ and $\boldsymbol{p}_j \in \mathcal{P}_i$ indicating to what degree $\boldsymbol{x}_i$ leaks personal information in $\boldsymbol{p}_j$.

We formulate the computation of alignment scores as an NLI problem. Namely, if $\boldsymbol{x}_i$ entails

Dataset	Year	Source	Human	Synthetic	Rewrite Type	Form of Private info	Size
NAP2 (Ours)	2025	PERSONACHAT	✓	✓	Delete / Obscure	privacy profile	Small
Self-Disclosure (Dou et al., 2024)	2024	Reddit Post	✗	✓	Obscure	text span	Small
SythPAI (Yukhymenko et al., 2024)	2024	Reddit sytle	✗	✗	✗	PII	Large
Text Anon. Benchmark (TAB) (Pilán et al., 2022)	2022	ECHR legal cases	✗	✗	Masking	PII	Large
TextWash(Kleinberg et al., 2022)	2022	Wikipedia Bio	✗	✗	Masking	PII	Medium

Table 2: Comparison of recent datasets for text anonymization or privacy-preserving rewriting. Human and Synthetic refers to if there is any generated rewrite for the dataset either from human or LLMs. Rewrite type includes deletion, obfuscation. Masking indicates that the detection private entities is redacted as their entity types.

$\boldsymbol{p}_j$, it is highly likely that $\boldsymbol{x}_i$ leaks information in $\boldsymbol{p}_j$. Specifically, we reuse the RoBERTa model trained on Multi-Genre Natural Language Inference (MNLI) corpus (Williams et al., 2018), which is available from Huggingface, to compute the probability of $p(y = \text{entail}|\boldsymbol{x}_i, \boldsymbol{p}_j)$ as s_{ij}. We find out that this simple approach significantly outperforms SPARSE-MAX and SHARP-MAX proposed in (Xu et al., 2020) on a random sample of 200 ground-truth pairs. We manually check the candidates among the pairs with a score higher than a threshold and keep only the well aligned ones.

For each selected sentence-persona pair, we recruit annotators from Amazon Mechanical Turk (AMT) to rewrite utterances w.r.t. the aligned persona sentences using both Deleting and Obscuring.

In our preliminary experiments, we observe that even though annotators endeavor to generate decent rewrites, many of them could not clearly identify and strictly stick to the required strategies. Therefore, we prepare a small sample of pairs as a pretest to select qualified annotators. In addition, we employ a rigorous procedure for quality check. We wrap up 15 sentence-persona pairs as a batch and ask annotators to rewrite them using the required strategies. Then, we manually check the rewritten batches, we only accept those that are written using the required strategy. The averaged acceptance rate of the rewrites is 47.97%, demonstrating the challenge of collecting a high-quality rewriting dataset with specific rewriting requirements. As a result, we collect 895 pairs annotated with one rewrite per strategy. We further split this corpus into a cross-validation (CV) set, a validation and a hold-out test set with 655, 140 and 100 instances, respectively.

Data Statistics. We analyze the manually curated corpus using averaged word length in sentences (Len.) and distinct unigrams divided by the total number of words (Dist.) (Li et al., 2016). The statistics of the dataset is given in Table 3. Deleting tends to produce more concise rewrites, while obscuring is slightly longer than ORIGINAL sentences. Although the average length increases, the diversity score for obscuring is still ascending, compared with original sentences. This shows the high diversity of word usage using obscuring.

	CV		Valid		Test	
	Len.	Dist.	Len.	Dist.	Len.	Dist.
ORI	13.7	0.148	13.6	0.257	13.5	0.248
DEL	8.0	0.190	8.4	0.298	8.5	0.279
OBS	14.1	0.160	13.9	0.266	14.3	0.250

Table 3: Statistics of original sentence (ORI), rewrites with deleting (DEL) and obscuring (OBS) on the CV set, validation and test set of the manually curated dataset, using average length (Len.) and distinct token (Dist.).

2.3 Synthetic Data Augmentation

We employ the RoBERTa NLI model to align utterances with persona sentences in the training set of PERSONA-CHAT and keep only the pairs with an entailment probability above 0.3. This threshold leads to high recall low precision alignments, so that GPT4 is employed to check if there is indeed a privacy leakage. Among them, we randomly sample 3900 pairs to generate synthetic rewrites by using GPT4. The resulting dataset is used to augment the training set of the manually created corpus to mitigate the data scarcity issue.

Prior studies show that GPT4 is one of the strongest few-shot learner (Brown et al., 2020). Therefore, we carefully design prompts and in-context examples to use for privacy-aware rewriting. Given an utterance-persona pair, we use the following prompt for a selected rewriting strategy.

> Rewrite this sentence, <deleting / obscuring> any private information.
> Example rewrites are:
> <IN-CONTEXT_EXAMPLES>
> Only return the rewritten sentence, nothing else.
> Private information present is: [$PERSONA].
> The sentence to rewrite is: [$UTTERANCE].

Here, $X denotes a placeholder for the corresponding information. The k in-context examples are

selected from a combination of the validation set of the manually curated corpus and a set of non-sensitive utterances which do not leak personal information. Each of the in-context examples in the validation set contains an utterance, a persona sentence, and a human rewrite using the given strategy, while an example from the non-sensitive set includes only an utterance. The in-context examples are found by k-nearest neighbour search using the sentence embeddings of utterances (Reimers and Gurevych, 2019). In this work, given an utterance, we select the top-1 most similar example from the validation set and one example from the non-sensitive set. The latter is used to instruct GPT4 that it should not rewrite an utterance if there is no privacy leakage detected.

2.4 Human Evaluation

Three university students are recruited to check their quality on a set of 100 instances sampled from the test set of the manual corpus. Hence, an utterance-persona pair in the sample includes a human rewrite, a rewrite from GPT-3.5 TURBO and GPT4 respectively. For each rewrite, a student is instructed to answer the following questions from the perspectives of privacy leakage (Q1), semantic relevance (Q2) and naturalness (Q3) which is detailed in Appendix A.1. Each question is answered by three university students. To deal with possible disagreements, we take the *majority vote* as the final answer. For annotation, the three annotators achieved Fleiss' Kappa(Falotico and Quatto, 2015) inter-annotator agreement score with 0.47 which is acceptable for classification problem. We further conducted closer examination with it. It revealed that this was largely due to disagreement from a single annotator. When isolating the annotations from the other two annotators, we observed substantially higher inter-annotator agreement, with Fleiss' Kappa values of 0.987 (Q1), 0.942 (Q2), and 0.883 (Q3). These figures indicate strong consistency between the two annotators and suggest that the lower overall scores were not the result of unclear guidelines or ambiguous questions, but rather individual annotator variability. We further incorporated the extra annotator to do the checks and showing consistent Kappa score with two students with 0.938(Q1), 0.913(Q2) and 0.864(Q3).

In order to use a score to summarize the performance w.r.t. each criteria, we calculate the percentage of choosing the option (a) as the majority vote for each question above on the human evaluation test set, referred to as SPRIVACY, SREL, and SNATURAL. They indicate the percentage of rewrites having no privacy leakage, complete semantic relevance, full naturalness, respectively.

	SPRIVACY	SREL	SNATURAL
Human_deleting	82.00%	76.00%	95.00%
GPT3.5_deleting	34.00%	94.00%	72.00%
GPT4_deleting	49.00%	92.00%	99.00%
Human_obscuring	81.00%	97.00%	98.00%
GPT3.5_obscuring	61.00%	90.00%	95.00%
GPT4_obscuring	66.00%	95.00%	99.00%

Table 4: Comparison between GPT-3.5 TURBO, GPT4, and human rewrites.

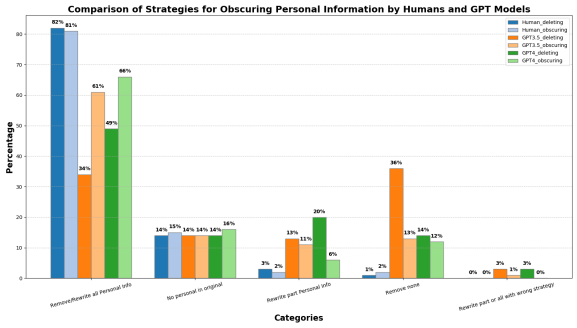

Figure 1: Human evaluation of privacy leakage.

To understand the quality of rewrites in our corpus, we compare GPT4 outputs with those of GPT-3.5 TURBO using the same prompts, as well as with human rewrites. The key results are summarized in Table 4. Human rewrites achieve the highest level of privacy protection with both strategies, outperform the best rewriting model GPT4 by at least 15%. Human rewrites with obscuring achieve the best balance between privacy and utility in comparison with alternative methods. Both OpenAI models completely preserve personal information in over 60% of utterances by using obscuring, but struggle to implement the deleting strategy for the same purpose. A close investigation on the percentages of individual Q1 answer in Fig. 1 demonstrates that both models fail to delete private expressions completely in over 34% of the utterances involving sensitive information. GPT-3.5 TURBO is significantly worse than GPT4 in terms of sanitization. Only a small proportion of the errors are attributed to applying an incorrect strategy.

3 Experiments

3.1 Rewriting Models

In this section, we establish a baseline approach to assess the efficacy of existing privacy protection solutions in removing private content within textual

messages. This evaluation is pivotal to addressing the critical question: "**Are current privacy protection solutions adequately equipped to conceal privacy-sensitive content in utterances?**"

For a comprehensive comparison of inference-time privacy preservation, we benchmark representative DP and rewriting approaches, including DPNR (Lyu et al., 2020), DP-Forward (Du et al., 2023), LLAMA-PARAPH, DP-PROMPT, FLAIR-SCRUBBING, and DP-BART. Briefly, DPNR injects Laplace noise into token representations, whereas DP-Forward perturbs embedding matrices to implement sentence-level local differential privacy. Detailed introduction and implementation specifics for all baselines are provided in App. A.4.

To quantify zero-shot rewriting capability, we evaluate the same pretrained LLMs under the prompt template from Sec. 2.3 without additional training, namely T5-BASE, LLAMA2-13B, GPT-3.5 TURBO, and GPT4. To distinguish zero-shot from supervised or baseline variants, we denote the former as T5_ZEROSHOT and LLAMA2-13B_ZEROSHOT. We only reported the GPT-4o as zero-shot method with automatic metrics to show the consistent performance for SOTA LLMs. For supervised baselines, we fine-tune T5-BASE on NAP^2 with and without GPT-4–based synthetic augmentation, and additionally fine-tune BART to assess dataset effectiveness independently of architecture. Finally, to examine training-time privacy defenses, we train T5–NAP^2 with and without DP-SGD (Abadi et al., 2016a) to characterize the impact of differential privacy on performance.

MIA setup. For downstream task privacy evaluation, we also would like to see if the schema of privacy rewrite can effectively mitigate the membership inference attack for original data as language model training is widely adapted for open domian corpus. Specially, we conducted a lightweight experiment by finetuning a compact target model (LLaMA3.2-1B) on human rewritten and machine rewritten sentences respectively to evaluate if the privacy information can be inferred from trained model using state-of-the-art self-prompt–calibrated membership inference attack SPV-MIA (Fu et al., 2024).

3.2 Evaluation Details

Prior studies focus on protect data privacy from membership inference attacks, reconstruction attacks, and sensitive attribute attacks etc. (Mattern et al., 2022). However, almost all of them focus on privacy preservation at the training time. In contrast, our target task is concerned with i) if a rewrite reveals personal information in a given persona, ii) preservation of non-sensitive content, and iii) naturalness of rewrites. Compared with the prior studies based on DP mechanisms, our setting is more close to that of natural language generation (NLG) tasks. Therefore, we evaluate the outcomes of the rewriting models by using NLG motivated automatic and human evaluation.

Method	PRIVACY_NLI	SPRIVACY	ROUGE-1	ROUGE-LSUM
DPNR	62.14%	25.00%	92.79%	92.79%
DP-Forward	36.42%	0.00%	99.91%	99.91%
DP-PROMPT	62.86%	0.00%	42.18%	41.89%
DP-BART	78.22%	1.00%	44.01%	43.15%
FLAIR-SCRUBBING	56.43%	0.00%	67.75%	67.89%
T5_ZEROSHOT-deleting	70.00%	10.00%	16.62%	12.61%
T5_ZEROSHOT-obscuring	45.00%	45.00%	29.58%	23.80%
LLAMA2-13B_ZEROSHOT-obscuring	79.28%	16.00%	40.86%	40.12%
LLAMA2-13B_ZEROSHOT-deleting	77.14%	14.00%	68.28%	67.53%
LLAMA-PARAPH-obscuring	82.86%	31.00%	21.72%	20.05%
LLAMA-PARAPH-deleting	76.42%	16.00%	56.29%	54.91%
GPT-3.5-obscuring	87.14%	61.00%	66.66%	65.76%
GPT-3.5-deleting	74.29%	34.00%	69.13%	68.48%
GPT-4-obscuring	92.14%	66.00%	73.24%	72.63%
GPT-4-deleting	90.00%	49.00%	77.48%	77.08%
GPT-4o-obscuring	84.29%	–	67.75%	67.24%
GPT-4o-deleting	89.29%	–	74.95%	74.92%
T5-NAP^2-GPT4	**93.81%**	**72.00%**	73.01%	72.78%

Table 5: Evaluation and comparison of baseline methods.

For human evaluation, we use the same questionnaires and the metrics introduced in Sec. 2.4 and ask annotators to answer each question in order to obtain the majority votes.

For all experiments involving model fine-tuning, we conduct five folds cross validation (CV) on the CV set of the manually curated corpus. In order to understand the usefulness of synthetic data, we also conduct experiments with the same models that augment the training set in each fold with 3,900 synthetic instances generated by GPT4.

3.2.1 Automatic Evaluation Metrics.

Privacy Leakage. We propose a novel metric, called PRIVACY_NLI, by using the ROBERTA model trained on the MNLI corpus, to infer to what degree it is possible to infer personal information in personas. As the NLI model classifies a pair of input texts into *entailed*, *contradicted*, or *neutral*, we adopt $P(\text{entailed}|\boldsymbol{x}, \boldsymbol{p})$ as the score of $privacy\_leakage$, e. Hence, we consider PRIVACY_NLI as 1- $privacy\_leakage$, denoting the privacy preserved by our method. The higher the metric, the more private information is preserved. To validate the effectiveness of this metric, we incorporated recent NLI models trained with variant MNLI corpus to show the consistency. Besides we also considered Sparse-MAX and Sharp-MAX as soft alignment score(Xu et al., 2020), the detailed

	SPRIVACY	SREL	SNATURAL
Human_deleting	**82.00%**	76.00%	95.00%
LLAMA2-13B _deleting	54.00%	49.00%	87.00%
T5-NAP^2-GPT4 _deleting	**72.00%**	91.00%	95.00 %
DPNR	1.00%	0.00%	19.00%
Human_obscuring	**81.00%**	97.00%	98.00%
DP-PROMPT	0.00%	1.00 %	0.00%
DP-BART	1.00%	10.00%	2.00%
FLAIR-SCRUBBING	0.00%	1.00%	0.00%
LLAMA2-13B _obscuring	12.00%	14.00%	86.00%
T5-NAP^2-GPT4 _obscuring	**53.00%**	93.00%	98.00%

Table 6: Human evaluation of the SOTA models.

	SNATURAL	LLM-NATURAL
Human_deleting	95.00%	4.14
LLAMA2-13B _deleting	87.00%	4.67
T5-NAP^2-GPT4 _deleting	95.00 %	4.44
DPNR	19.00%	2.01
Human_obscuring	98.00%	4.37
DP-PROMPT	0.00%	1.14
DP-BART	2.00%	1.71
FLAIR-SCRUBBING	0.00%	3.05
LLAMA2-13B _obscuring	86.00%	4.85
T5-NAP^2-GPT4 _obscuring	98.00%	4.36

Table 7: LLM as Naturalness Judge compared with Human preference.

comparsion can be found in Appendix. A.7.

Semantic Relevance. For assessing the preservation of semantic content, we consider ROUGE-1 and ROUGE-LSUM (Lin, 2004) to compare generated rewrites with the corresponding references detailed introduction can be found in Appendix. A.4.

3.3 Results and Discussions

Efficacy of NAP^2. Table 5 reports the evaluation of all methods. T5-BASE fine tuned on the human rewrites and the synthetic data using both strategies outperform the DP based methods and zero-shot LLMs by a wide margin. DPNR preserves more privacy than DP-Forward, but results in a dramatic drop of information utility. The generated texts often have completely different meanings and have substantial grammatical errors, though some of them are still fluent. In contrast, DP-Forward mostly copies inputs to outputs but rarely hide sensitive information. LLAMA-PARAPH produces frequently irrelevant texts, hence have fairly low ROUGE-1 and ROUGE-LSUM scores. Besides, for convention personally identifiable information scrubbing method FLAIR-SCRUBBING, it can not effectively remove the private information in open-ended domain, only 40.71% examples are successfully removing PII tokens. For DP-PROMPT and DP-BART, even PRIVACY_NLI are outperformed than other baseline models, the paraphrasing impairs the semantic of original sentence leading to low ROUGE-1 score.

We further investigate the rewriting quality w.r.t. each strategy based on human evaluation. We use the T5-BASE model trained on the human rewrites and the synthetic data with both strategies, and apply it on the hold-out test set of each strategy. Table 6 shows that the T5-BASE model achieves superior performance over the baselines with both strategies. The naturalness of all generated rewrites is on par with that of human rewrites. Both zero-shot LLAMA2-13B models perform better than the best DP method DPNR, which mostly perturbs non-sensitive contents or yields repeated words. The near-zero SPRIVACY scores observed for these methods stem from the nature of noise injection in embeddings or numeric representations. This results in two extremes: either no change in output due to insufficient perturbation or complete distortion of the generated output. The overall results are encouraging for a wide range of applications on edge devices, because our corpus is not huge and T5-BASE contains only a few million parameters, which is a few hundred times smaller than LLAMA2-13B, GPT-3.5 TURBO and GPT4 and GPT4O.

Alignments between Automatic metrics and Human Evaluation. We compare the ranking using PRIVACY_NLI with the corresponding human judgments in Table 5. T5-NAP^2-GPT4 obtains the highest 1-PRIVACY_NLI of 93.81% in automatic evaluation, matching the highest SPRIVACY with 72.00%. The results are aligned well among the rewriting models using the obscuring. However, PRIVACY_NLI does not rank all rewriting models using deleting in the same manner as humans. To quantify the alignments, we calculate a Spearman's correlation of 0.70 between PRIVACY_NLI and SPRIVACY among all models to demonstrate PRIVACY_NLI. The correlation between the models using obscuring reaches even 0.83.

Naturalness Assessment via LLM-as-judge We also consider LLM as judge to score the naturalness of generated sentence. Specifically, we reuse the question from human questionnaire about naturalness and convert it to prompt template with score scale from 1-5. The prompt can be found in the Appendix A.9 The results in Table 7 demonstrate a generally strong alignment between **LLM-judged naturalness** and **human preference** (SNATURAL), particularly for high-performing models such as LLAMA2-13B _deleting, T5-NAP^2-GPT4 _deleting, and T5-NAP^2-GPT4 _obscuring, where LLM

	SPRIVACY	SREL	SNATURAL
Human_deleting	**82.00%**	76.00%	95.00%
T5-NAP^2-GPT4 _deleting	72.00%	91.00%	95.00%
non-Syn_deleting	65.00%	92.00%	93.00%
Human_obscuring	**81.00%**	97.00%	98.00%
T5-NAP^2-GPT4 _obscuring	53.00%	93.00%	98.00%
non-Syn_obscuring	4.00%	92.00%	93.00%

Table 8: Human evaluation results with and without synthetic data.

PRIVACY_NLI	Mixed corpus	MNLI	SPRIVACY
DPNR	80.54%	53.69%	25%
DP-Forward	68.87%	48.06%	0%
DP-PROMPT	71.85%	82.76%	0%
DP-BART	85.87%	80.72%	1%
FLAIR-SCRUBBING	62.16%	61.49%	0%
T5_ZEROSHOT-deleting	61.33%	49.14%	10%
T5_ZEROSHOT-obscuring	61.33%	19.26%	45%
LLAMA2-13B_ZEROSHOT-obscuring	81.85%	45.23%	16%
LLAMA2-13B_ZEROSHOT-deleting	87.04%	76.68%	14%
LLAMA-PARAPH-deleting	79.42%	65.91%	31%
LLAMA-PARAPH-obscuring	82.69%	52.21%	16%
GPT-3.5-obscuring	90.33%	53.51%	61%
GPT-3.5-deleting	61.33%	49.60%	34%
GPT-4-obscuring	95.36%	56.50%	66%
GPT-4-deleting	89.26%	47.59%	49%
Human Rewrite	97.01%	63.20%	82%

Table 9: PRIVACY_NLI score with DEBERTA as backbone finetuned with mixed entailment corpus and MNLI.

scores ($\geq$ 4.36) closely reflect human-rated naturalness ($\geq$ 95%). However, notable discrepancies emerge for low-performing systems like DP-PROMPT and FLAIR-SCRUBBING, where LLMs assign moderately high naturalness scores (e.g., 3.05 for FLAIR-SCRUBBING) despite near-zero human ratings. This indicates that while LLMs can approximate human judgments in many cases, they may overestimate the fluency or coherence of outputs that humans find unnatural, underscoring the need for further calibration of LLM-based evaluation frameworks.

Usefulness of the Synthetic Data. Table 8 shows the result of using synthetic data for training rewriting models. We compare two different strategies: deleting and obscuring. The results shows that the model performs better with the synthetic data for both tasks. In particular, the model preserves more non-personal information compared to human rewrites in the deleting task. With the synthetic data training the models, the model performance is 7% better than the non-synthetic data model in terms of deleting. The biggest gain of the synthetic data is obtained for improving the privacy protection of the rewriting model using obscuring.

NLI score with different backbones In this section, we present the PRIVACY_NLI scores obtained using the DEBERTA model as the backbone, fine-tuned with different corpora, including a mixed entailment corpus and MNLI, to evaluate alignment with SPRIVACY, which serves as our human evaluation metric. As shown in Table 9, the DEBERTA model achieves varying levels of performance across datasets. Models trained on the mixed corpus, such as DP-BART and -deleting, achieve PRIVACY_NLI scores of 85.87% and 95.36%, respectively, with the former reaching 1% and the latter 66% alignment with SPRIVACY. Notably, human rewrites achieve a PRIVACY_NLI score of 97.01%, with an alignment of 82% with SPRIVACY. This comparison underscores the capability of different backbones and training strategies to achieve results close to human-level performance about privacy preservation while maintaining alignment with human evaluation metrics.

Method	**AUC**	**ASR**	**TPR@1%FPR**
Human Rewrite	0.68	0.665	0.15
T5–NAP^2–GPT4	0.51	0.58	0.01

Table 10: MIA results on a compact target model (higher = worse privacy). Rewrites from our trained model exhibit near-random attackability.

3.4 Privacy auditing via Membership Inference (MIA)

To test the privacy implication of privacy rewrite, we employ model finetuning as the targeted downstream task. In this case, membership inference attack is widely adopted to measure privacy leakage via training time memorization of model. By applying the privacy rewrite, we would to reduce the memorization of original context via privacy rewrite. We employ the self-prompt–calibrated membership inference attacks(SPV-MIA)(Fu et al., 2024) as our attack method as it follows the practical attack scenario which does not requires obtaining real datasets. We fine-tune a LLaMA3.2-1B target model on human rewritten and model rewrites respectively to measure attack power and comparing it with the model finetuned using the original dataset via AUC, ASR, and TPR@1%FPR.

Findings. Attacks on our *model rewrites* are near random (AUC $\approx$ 0.51), indicating low memorization. The *human rewrites*—which intentionally preserve more non-private content—show higher TPR@1%FPR, consistent with their stronger utility. This complements our main results: explicit deletion/obscuring of private attributes reduces downstream attackability while keeping semantics usable.

4 Related Work

The field of controllable text style transfer focuses on modifying specific attributes in texts, such as formality (Briakou et al., 2021) and sentiment (Li et al., 2018a, 2022) while preserving the core semantic content. The advancement of text rewriting tasks is heavily dependent on the availability of high-quality corpora to assess generation quality. Rao and Tetreault (2018) collected a large-scale corpus GYAFC for initiating the research of formality style transfer to rewrite formal language. As for our task sensitive to privacy, which demands sophisticated alignment in rewriting utterances, the construction of a specialized corpus for high-quality privacy-sensitive rewrites are crucial.

There is a growing interest in protecting user privacy (Chen et al., 2020; Tigunova et al., 2019; Xu et al., 2019; Bevendorff et al., 2019) in NLP tasks. One way of protecting privacy is to implicitly remove the information in decision models, for example perturbing the representations via adversarial training (Li et al., 2018b; Elazar and Goldberg, 2018; Barrett et al., 2019) or differential privacy (Fernandes et al., 2019; Bo et al., 2019). In text rewriting which is close to our rewriting approach, local differential privacy are recently adapted to protect the data by adding customized noise (Igamberdiev et al., 2022b; Igamberdiev and Habernal, 2023). Such adaptations in rewriting system mitigate the privacy leakage risk of original input however result in complete semantic change of inputs as the noise is independently drawn from the data and task. We consider a more generalised rewriting setting where the naturalness and general meaning of sentence are preserved.

Another series of work suggested to generate new sentences with less sensitive information (Emmery et al., 2018; Xu et al., 2019). Recent work has explored prompting LLMs to rewrite sentences containing private information, aiming to obscure sensitive content (Emmery et al., 2018; Xu et al., 2019; Staab et al., 2024). However, these approaches often rely on LLMs' internal knowledge and struggle to align with nuanced human privacy expectations (Dou et al., 2024). However in these works, the author does not control how private information to be rewritten in explicit way which may weaken the control for required privacy and human preference. Even advanced techniques like self-disclosure abstraction or adversarial anonymization face challenges in achieving robust, user-aligned privacy and often depend on powerful cloud-based models. In contrast, our work study private rewrite via more diversified free text and supports two rewriting strategies, offering a more flexible and general setting (Strengers et al., 2020).

5 Conclusion

We introduce *naturalness* to the tasks of privacy-preserving text rewriting and collect a corpus NAP^2 based on PERSONA-CHAT. The fundamental concept involves training models to learn human strategies, namely deleting and obscuring, for inference-time privacy. The T5-BASE model trained on our corpus outperforms competitive zero-shot LLMs and DP methods by a wide margin. This work paves the way for future research on LLM-based rewriting techniques with a new focus on preserving naturalness of rewriting.

Ethical Statement

In this paper, we align our research practices with the principles outlined in the ACL Code of Ethics, fully endorsing its values. Our investigation has been conducted in compliance with these ethical standards.

The creation and assessment of NAP^2 have been conducted with a keen awareness of ethical considerations, especially regarding the involvement of human annotators. The necessity for human-annotated data to train conditional independence classifiers in our method is recognized as demanding significant effort. We have taken careful measures to ensure that this process is ethically sound, honoring the annotators' contributions by respecting their time and providing equitable compensation.

Moreover, the central objective of NAP^2 is to assess the relevance of generated responses in relation to their persona information and the difference between human evaluation and proposed automated metrics. The system is engineered to assign scores on a continuous scale from 0 to 1, with higher scores denoting greater relevance. It is designed to yield only these scores, without generating any information that could be deemed harmful or violate privacy.

Limitation

Due to budgetary constraints associated with this project, we were unable to engage a vast number of annotators to rewrite the extensive dialogue

datasets with respective rewrite strategies. Consequently, NAP^2 we compiled is somewhat limited in scope. While NAP^2 possesses sufficient volume to validate the core assertions of our study, it might not fulfill the expansive needs of commercial deployments. Industrial entities interested in utilizing our dataset could potentially address this limitation by adopting prompt tuning techniques or employing additional annotators to expand the dataset in accordance with our outlined methodology.

Our evaluation metric is specifically designed to assess the relevance of the generated responses. Although it demonstrates superior performance over baseline metrics in terms of privacy preservation and naturalness, the advantage it presents in relevance and specificity is less pronounced. Therefore, the development of innovative metrics tailored to specific evaluation criteria presents a valuable avenue for our future research endeavors.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016a. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016b. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6331–6336.

Janek Bevendorff, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Heuristic authorship obfuscation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1098–1108.

Haohan Bo, Steven HH Ding, Benjamin Fung, and Farkhund Iqbal. 2019. Er-ae: differentially-private text generation for authorship anonymization. *arXiv preprint arXiv:1907.08736*.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! xformal: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Xiaolin Chen, Xuemeng Song, Ruiyang Ren, Lei Zhu, Zhiyong Cheng, and Liqiang Nie. 2020. Fine-grained privacy detection with graph-regularized hierarchical attentive representation learning. *ACM Transactions on Information Systems (TOIS)*, 38(4):1–26.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2024. Reducing privacy risks in online self-disclosures with language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13732–13754.

Minxin Du, Xiang Yue, Sherman SM Chow, Tianhao Wang, Chenyu Huang, and Huan Sun. 2023. Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 2665–2679.

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

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21.

Chris Emmery, Enrique Manjavacas, and Grzegorz Chrupała. 2018. Style obfuscation by invariance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 984–996.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49:463–470.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *International Conference on Principles of Security and Trust*, pages 123–148. Springer, Cham.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. 2024. Membership inference attacks against fine-tuned large language models via self-prompt calibration. *Advances in Neural Information Processing Systems*, 37:134981–135010.

Timour Igamberdiev, Thomas Arnold, and Ivan Habernal. 2022a. Dp-rewrite: Towards reproducibility

and transparency in differentially private text rewriting. In *Proceedings of the 29th International Conference on Computational Linguistics*, page (to appear), Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Timour Igamberdiev and Ivan Habernal. 2023. Dp-bart for privatized text rewriting under local differential privacy. *arXiv preprint arXiv:2302.07636*.

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

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018a. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018b. Towards robust and privacy-preserving text representations. *arXiv preprint arXiv:1805.06093*.

Zhuang Li, Lizhen Qu, Qiongkai Xu, Tongtong Wu, Tianyang Zhan, and Gholamreza Haffari. 2022. Variational autoencoder with disentanglement priors for low-resource task-specific natural language generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10335–10356.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363. IEEE.

Lingjuan Lyu, Xuanli He, and Yitong Li. 2020. Differentially private representation for nlp: Formal guarantee and an empirical study on privacy and fairness. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2355–2365.

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.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884*.

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. IEEE.

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

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*, 21:1–67.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

David Sánchez, Montserrat Batet, and Alexandre Viejo. 2014. Utility-preserving privacy protection of textual healthcare documents. *Journal of biomedical informatics*, 52:189–198.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

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

Yolande Strengers, Lizhen Qu, Qiongkai Xu, and Jarrod Knibbe. 2020. Adhering, steering, and queering: Treatment of gender in natural language generation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Latanya Sweeney. 2002. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):571–588.

Anna Tigunova, Andrew Yates, Paramita Mirza, and Gerhard Weikum. 2019. Listening between the lines: Learning personal attributes from conversations. In *The World Wide Web Conference*, pages 1818–1828.

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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Qiongkai Xu, Lizhen Qu, Zeyu Gao, and Gholamreza Haffari. 2020. Personal information leakage detection in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6567–6580.

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.

Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin Vechev. 2024. A synthetic dataset for personal attribute inference. *Advances in Neural Information Processing Systems*, 37:120735–120779.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

A Appendix

A.1 Question design for human evaluation

Q1: The rewrite deletes/obfuscates _ ?

(a) None of the key information in the personal information and the original utterance does contain personal information.

(b) None of the key information in the personal information, because the original utterance does not contain personal information.

(c) At least one key information in the personal information (if the rewrite uses both correct and incorrect strategies, only evaluate the part that uses the correct strategy).

(d) All key information in the personal information (using the correct strategy only).

(e) At least one or all key information in the personal information (using the incorrect strategies only).

Q2: The rewrite _.

(a) Accurately preserves the meaning of the original sentence.

(b) Basically the same meaning but does not cover some minor content.

(c) Has a minor resemblance to the meaning of the original sentence, however, it is also misleading.

(d) Empty sentence or does not reflect the meaning of the original sentence at all.

Q3: The rewrite is able to retain _ in the original utterance that is not covered in the personal information.

(a) has no grammatical mistakes and the sentence is coherent.

(b) has some grammatical mistakes and the sentence is less coherent

(c) is full of grammatical mistakes and the sentence is not coherent

A.2 PERSONA-CHAT

The PERSONA-CHAT dataset (Zhang et al., 2018) is a crowd-sourced corpus designed to facilitate research in personalized open-domain dialogue systems. Each conversation in the dataset involves two speakers, each assigned a distinct persona comprising 4–5 profile sentences. These personas guide the dialogue, encouraging participants to engage in conversations that reflect their assigned characteristics.

The dataset encompasses:

- **1,155** unique personas, each with at least 5 profile sentences.
- **10,907** dialogues totaling over **162,000** utterances.
- A division into training, validation, and test sets, with 100 personas reserved for validation and another 100 for testing.

This structure promotes the development of dialogue agents capable of maintaining consistent and engaging personalities throughout interactions.

A.3 Prompt template for synethetic Data

The prompt template used across the paper is shown as 2. We use three nearest examples drawn from the training set as prompting example. Each example contains two cases if the raw persona information is provided. And objective for the prompt is to rewrite given sentence with specified strategy.

Example 1:
If I ask you to rewrite [example #1]
containing personal information [persona #1]
by <deleting/obscuring> private information, you should return [target #1]
If I ask you to rewrite [example #1]
containing personal information [empty]
by <deleting/obscuring> private information, you should return [example #1]

Example 2:
If I ask you to rewrite [example #2]
containing personal information [persona #2]
by <deleting/obscuring> private information, you should return [target #2]
If I ask you to rewrite [example #2]
containing personal information [empty]
by <deleting/obscuring> private information, you should return [example #2]

Example 3:
If I ask you to rewrite [example #3]
containing personal information [persona #3]
by <deleting/obscuring> private information, you should return [target #3]
If I ask you to rewrite [example #3]
containing personal information [empty]
by <deleting/obscuring> private information, you should return [example #3]

Rewrite this sentence, deleting any private information.
Only return the rewritten sentence, nothing else.
Private information present is: [input persona]
Sentence to rewrite: [input utterance]

Figure 2: Prompt template for T5-NAP^2

A.4 Experiment Details

A.4.1 Evaluation metrics

Details of the evaluation metrics for semantic relevance are provided below.

- ROUGE-1 (Lin, 2004): It is a widely used evaluation metric measuring the overlap of unigrams between a generated text and a set of references.
- ROUGE-LSUM: It is a variant of ROUGE-L, tailored to evaluate longer texts by summarizing the longest common sub-sequences between an output text and a set of references.

A.4.2 Baseline Methods

DPNR. It stands for Differentially Private Neural Representation, which applies Laplace noise to distributed representations of words in order to randomly drop sensitive words or replace sensitive words with non-sensitive ones. We compare the cosine similarity between each word in an input utterance with those in the corresponding persona, and pick the top-k most similar ones.

DP-Forward. This method perturbs embedding matrices and multi-head attention layers during each forward pass of a language models by achieving a sentence level LDP. When adapting this approach to T5-BASE for inference, we mainly perturb embedding matrices, because the DP mechanism for attention layers is mostly useful for protecting privacy at the training time.

LLAMA-PARAPH. Mattern et al. (2022) points out the limitations of word-level LDP and propose to paraphrase input texts with lower temperature to achieve a sentence-level LDP. We implement this approach by using LLAMA2-13B.

DP-PROMPT. Utpala et al. (2023) utilizes zero-shot prompting and large language model to generate document paraphrasing to prevent author de-

anonymization attack which comprise the privacy of text owner with predefined utility constrain.

DP-BART. The method is a privatized text rewriting system incorporates LDP. The system leverages the LPD paradigm to perform model rewriting using BART model to protect input data which tackles same challenge like us.

FLAIR-SCRUBBING. we also adapt the scrubbing method used in (Lukas et al., 2023) as our baseline. We set FLAIR-SCRUBBING as baseline to test if the automatic method can effectively remove private information from sentences.

Zero-Shot LLMs. To compare with the LLMs fine-tuned on our corpus, we apply the same prompts to the same pre-trained LLMs without any training. Specifically, we consider T5-BASE, LLAMA2-13B, GPT-3.5 TURBO and GPT4 and apply the prompt template introduced in Sec. 2.3. To distinguish from the fine-tuned models, the T5-BASE and LLAMA2-13B in the zero-shot setting is referred to as T5_ZEROSHOT and LLAMA2-13B_ZEROSHOT, respectively.

T5-NAP^2. By using the same prompts as the zero-shot version, we fine tune T5-BASE on the training set of the manually curated corpus, with or without augmenting them with synthetic data. The prompts are similar to those used by zero-shot models detailed in A.3.

T5-NAP^2-DP. To simulate the use cases that the training data of the rewriting models contains sensitive information, we apply DP-SGD (Abadi et al., 2016a) when fine-tuning the T5-BASE model in order to understand to what degree the DP mechanism impacts the inference quality of the rewriting models and shed light on future research directions.

A.5 Implementation Details

In our experiment, we consider T5-BASE as our targeted rewrite model, we set optimal hyperparameters for model fine tuning with learning rate of $5e^{-4}$ and beam search as decoding method with generative temperature of 0.2. In the model fine-tuning, we set noise multiplier of DP-SGD (Abadi et al., 2016b) to 0.001 to gain minimal influence for model result. In baseline experiments, for two DP methods applied to echo language model, we consider the empirically optimal noise multipliers 0.01 and epsilon to 3 with one word masked for DPNR. As for DP-Forward-utility, we set the key noise hyperparameters delta to $1e^{-5}$ and epsilon at 7 to obtain the impact with small noise gap, while for DP-Forward-privacy, we set the hyperparameters to $2e^{-5}$ and 8 for delta and epsilon respectively. The remaining hyperparemeters are the same as with the ones reported in the corresponding papers.

A.6 Impact of DP-SGD.

Table 11 shows results of models trained with and without DP-SGD. The purpose is to understand to what degree the widely used DP method can influence rewriting quality if the training data is sensitive. Comparing these two settings with human rewrites, there is a slight performance drop of around 3% with DP-SGD. However, DP-SGD provides a privacy guarantee during training which is useful when the training data is sensitive. When comparing with automatic metrics, as shown in Table 12, there is only a 1% performance drop in terms of privacy leakage if DP-SGD is applied. For preservation of semantic contents, MAUVE scores show little differences between using and not using DP-SGD, meaning our proposed rewriting approaches are compatible with the DP based training algorithms for more sensitive scenarios.

	SPRIVACY	SREL	SNATURAL
Human_deleting	**82.00%**	76.00%	95.00%
DP_deleting	59.00%	88.00%	99.00%
non-DP_deleting	63.00%	82.00%	96.00%
Human_obscuring	**81.00%**	97.00%	98.00%
DP_obscuring	29.00%	90.00%	98.00%
non-DP_obscuring	32.00%	88.00%	93.00%

Table 11: Human evaluation results with and without DP-SGD.

A.7 Private Information Alignment

Three alignment techniques were evaluated to determine their effectiveness in identifying utterance-persona associations: the RoBERTa MNLI entailment model, Sparse-MAX, and Sharp-MAX. The latter two algorithms, originally proposed for token-level alignment, compute an alignment matrix where each entry represents the probability that a token in an utterance leaks information about a token in a persona. Specifically, for a token i in the utterance and a token j in the persona, the alignment score in row i, column j denotes the leakage likelihood.

Since our task requires sentence-level alignment rather than token-wise alignment. We modified algorithms to compute sentence-level alignment probabilities over a sample of 200 utterance-persona pairs. The goal was to correctly identify the persona associated with each utterance based

DP	Real	Synth	LLM	PRIVACY_NLI	ROUGE-1	ROUGE-LSUM
False	1300	0	-	0.9190 ± 0.1077	0.6946	0.6924
False	1300	3900	GPT-3	0.9174 ± 0.0903	0.7143	0.7122
False	1300	3900	GPT-4	$\mathbf{0.9381 \pm 0.0870}$	0.7301	0.7278
True	1300	0	-	0.9398 ± 0.0759	0.7338	0.7316
True	1300	3900	GPT-3	0.9243 ± 0.0908	0.7368	0.7351
True	1300	3900	GPT-4	0.9297 ± 0.1135	0.7446	0.7428

Table 12: Evaluation for DP and combination of synthetic data and human rewrites.

on alignment scores, such that the correct utterance-persona pair receives a high score while unrelated pairs do not.

To make binary alignment decisions, a fixed threshold was applied to the alignment scores. Probabilities below the threshold were interpreted as indicating no alignment, and those above it as indicating alignment. An ideal model would achieve perfect alignment performance, with both precision and recall equal to 1. Our empirical analysis revealed that Sparse-MAX and Sharp-MAX did not generalize well to the sentence-level alignment scenario, as shown in Table. 13. This result is unsurprising given that these algorithms were originally designed for fine-grained, token-level applications. Furthermore, their performance may have been affected by the lack of hyperparameter optimization specific to the sentence-level setting. Due to time constraints, a comprehensive exploration of these configurations was not feasible. Nevertheless, future work may revisit these methods as viable options, contingent on further tuning.

In contrast, the RoBERTa-based MNLI entailment model demonstrated strong alignment performance and required minimal adaptation. Following empirical threshold analysis, a decision boundary of 0.3 was selected for determining alignment. This threshold yielded favorable results and served as the primary alignment mechanism in subsequent experiments.

A.8 Limitation of LDP at Inference Time

Typical scenarios for privacy protection at inference time include i) dataset release; ii) sending queries involving sensitive queries to LLMs hosted on untrusted servers. Local DP can be one possible solution to adding noise locally for individual data releasing and it have more relaxed definition for user input. LDP is designed to make local data pairs indistinguishable and work generally on a sample of instances. The mainstream LDP methods add random noise to local examples to balance privacy and utilities. The collection of modified instances are aggregated to obtain certain statistics for target tasks. The aggregation step is important to mitigate the negative effects of noise for information utility. However, privacy protection at inference time does not allow any aggregation operation among a set of instances and requires finding a tradeoff between utility and privacy for individuals. Thus LDP based text rewriting methods either add too much noise to destroy the utility of information or retain original content involving sensitive information. Our experiments demonstrate the SOTA methods based on LDP empirically and show the promising research direction using our dataset.

A.9 Naturalness Judgment Template

We reuse the questionnaire question of naturalness to form the naturalness template for GPT-4o. We rescale the naturalness from 1-5 where 1 means very unnatural and 5 means perfectly natural. We prompt model to generate a JSON like result to score the rewritten sentence and provide the explanation on it. The detailed template is shown Figure. 3

Utility test for downstream task To further evaluate the utility of rewrites for LLMs, we conducted additional experiments to compare the LLM's responses generated by original inputs and their rewrites as the downstream task. The original PERSONA-CHAT is designed in a multiple round and chit-chat manner. We locate the input utterance in our datasets in the position of dialogue and compare the generated response with original response which can be formed as the response generation task. Specifically, we feed original texts and their rewrites respectively to the llama3.2-3B-Instruct and compare their responses with candidate responses collected from PersonaChat dataset, from which we sampled our dataset. The candidate set contains both ground-truth and implausible responses written by human.

Generated responses were ranked by calculat-

Table 13: Private information alignment results.

Family	Threshold	Recall	Precision	Min	Max	Mean	Frobenius Norm	1-Norm
RoBERTa Entailment	0.20	0.71	0.25	0.00	0.99	0.04	12.81	8.08
	0.25	0.69	0.27	0.00	0.99	0.04	12.81	8.08
	0.30	**0.69**	**0.28**	**0.00**	**0.99**	**0.04**	**12.81**	**8.08**
	0.35	0.68	0.29	0.00	0.99	0.04	12.81	8.08
	0.40	0.68	0.29	0.00	0.99	0.04	12.81	8.08
	0.80	0.57	0.35	0.00	0.99	0.04	12.81	8.08
Sparse-MAX	0.20	1.00	0.01	1.00	1.00	1.00	99.50	99.00
	0.25	1.00	0.01	1.00	1.00	1.00	99.50	99.00
	0.30	1.00	0.01	1.00	1.00	1.00	99.50	99.00
	0.35	1.00	0.01	1.00	1.00	1.00	99.50	99.00
	0.40	1.00	0.01	1.00	1.00	1.00	99.50	99.00
	0.80	1.00	0.01	1.00	1.00	1.00	99.50	99.00
Sharp-MAX	0.20	1.00	0.01	0.30	0.43	0.35	35.77	38.60
	0.25	1.00	0.01	0.30	0.43	0.35	35.77	38.60
	0.30	1.00	0.01	0.30	0.43	0.35	35.77	38.60
	0.35	0.64	0.01	0.30	0.43	0.35	35.77	38.60
	0.40	0.02	0.03	0.30	0.43	0.35	35.77	38.60
	0.80	0.00	NaN	0.30	0.43	0.35	35.77	38.60

ing the cosine similarity of their embeddings to those of ground-truth responses, using a Sentence-BERT(Reimers and Gurevych, 2019). As shown in Table. 15, the response similarity between the original inputs and their rewrites is 35.99%. Both the responses generated from the original inputs and those from the rewrites achieve comparable similarity scores to the ground truth. It is also worth noting that the responses generated from the rewrites are closer to the ground truth than the implausible ones. Therefore, our rewrites achieve similar utility as original inputs.

Generalization of Datasets To further evaluate the generalization ability of our dataset, we consider the rewriting model BART as our backbone and fine-tune the model on our datasets. BART is an efficient model compared to state-of-the-art language models, with only 175 million parameters. As shown in Table 14, the results demonstrate that a generative model with a small parameter size can still effectively adopt human rewrites from our dataset. The PRIVACY_NLI score is 85.74%, which is competitive with GPT-3.5, and the ROUGE-LSUM score is 57.32%, indicating higher consistency in language compared to LLAMA-PARAPH. Considering the gap in parameter size, the results show that our dataset can enable models like BART and T5 to achieve competitive performance in human-like rewriting tasks.

	PRIVACY_NLI	ROUGE-1	ROUGE-LSUM
GPT-3.5-deleting	74.29%	69.13%	68.48%
LLAMA-PARAPH-deleting	76.42%	56.29%	54.91%
BART-NAP^2	85.74%	59.435	57.32%
T5-NAP^2-NAP^2	93.81%	73.01%	72.78%

Table 14: NAP^2 finetuned BART with baseline method.

	Original	Responses with Rewrites	Implausible Responses
Ground Truth Responses	82.00%	76.00%	95.00%
Rewrite	34.00%	94.00%	72.00%

Table 15: Response similarity Comparison between varying types of inputs.

LLM Naturalness Judgment Prompt

You are an expert linguist. Your task is to assess the naturalness of a given sentence — how fluent, human-like, and typical it sounds in everyday language use.

Rate the sentence on a scale from 1 to 5:

- **1** = Very unnatural (full of grammatical mistakes, incoherent)
- **2** = Mostly unnatural
- **3** = Somewhat natural (acceptable, but some issues)
- **4** = Mostly natural (minor issues)
- **5** = Very natural (fluent, coherent, no errors)

Sentence: `"Sentence to Assess"`

Only provide the score and a brief explanation in the following JSON format:

```
{"score": X, "explanation": "..."}
```

Figure 3: Prompt used for LLM-based naturalness judgment.