reteLLMe: Design Rules for using Large Language Models to Protect the Privacy of Individuals in their Textual Contributions

Mariem Brahem[1,2], Jasmine Watissee[3,1], Cédric Eichler[4,1], Adrien Boiret[4,1], Nicolas Anciaux[1,2], and Jose Maria de Fuentes[5,1]

[1] Inria, Petscraft project-team, <firstname.lastname@inria.fr>
[2] Université Paris Saclay - Versailles, <firstname.lastname@uvsq.fr>
[3] Institut Polytechnique de Paris, <firstname.lastname@polytechnique.edu>
[4] INSA Centre Val de Loire, <firstname.lastname@insa-cvl.fr>
[5] Universidad Carlos III de Madrid, <josemaria.defuentes@uc3m.es>

Abstract. The advanced inference capabilities of Large Language Models (LLMs) pose a significant threat to the privacy of individuals by enabling third parties to accurately infer certain personal attributes (such as gender, age, location, religion, and political opinions) from their writings. Paradoxically, LLMs can also be used to protect individuals by helping them to modify their textual output from certain unwanted inferences, opening the way to new tools. Examples include sanitising online reviews (e.g., of hotels, movies), or sanitising CVs and cover letters. However, how can we avoid miss estimating the risks of inference for LLM-based text sanitisers? Can the protection offered be overestimated? Is the original purpose of the produced text preserved?
To the best of authors knowledge, no previous work has tackled these questions. Thus, in this paper four design rules (collectively referred to as *reteLLMe*) are proposed to minimise these potential issues. We validate these rules and quantify the benefits obtained in a given use case – sanitising hotel reviews. We show that up to 76% of at-risk texts are not flagged as such without fine-tuning. Moreover, classic techniques such as BLEU and ROUGE are shown to be incapable of assessing the amount of purposeful information in a text. Finally, a sanitisation tool based on *reteLLMe* demonstrates superior performance to a state-of-the-art sanitiser, with better results on up to 90% of texts.

Keywords: LLM · Privacy · Inference · Anonymisation

1 Introduction

Large Language Models (LLM) and related generative artificial intelligence techniques are on the rise. They are able to perform complex tasks such as video generation or speech synthesis, to name a few [2].

Despite their countless advantages, LLMs pose a serious threat to privacy by means of inferences [19,5]. Indeed, their ability to accurately predict sensitive

attributes (such as age, gender or political beliefs) of the author of a text that is believed to be benign has already been demonstrated [16]. This may lead to the complete deanonymization of the author of a piece of work that was intended to remain unknown, thus leading to undesired consequences.

Conversely, the very same inference capability of LLMs can be applied to mitigate the privacy threat. Previous works have already shown how LLMs can be converted into large-scale anonymizers, thus transforming a piece of data (say text, video, image, etc.) into another one that can reduce the precision of LLM-based inferences [17,18]. Some approaches have also considered the utility of the generated texts as a feature to preserve [4].

There are a number of motivating use cases for such a privacy-preserving use of LLMs. For example, in the context of online reviews, whether for hotels, products, or services, it is essential to protect the anonymity of reviewers to encourage honest and unbiased feedback. Here, a LLM can assist by sanitizing text to conceal Personally Identifiable Information (PII), ensuring consumer opinions are shared without fear of personal exposure. Similarly, in professional environments, such as job applications, ensuring that work-related documents such as reports or cover letters may not leak any highly sensitive information such as religious beliefs or political thoughts could help mitigate discrimination.

A great number of recent efforts revolve around using LLMs as privacy-preserving tools, such as [14,17], to name a few. Although they exhibit promising performance features, their use of LLMs does not follow any particular design criteria. This leads to undesired effects in text utility, user privacy or both. Therefore, it is essential to build those privacy-preserving LLMs in a sound manner, providing reliable evidence of their effectiveness.

To address these issues, this paper proposes a set of design rules (collectively referred to as $reteLLMe$) to build privacy-preserving LLMs and evaluate their efficiency. They provide solid grounds to achieve an optimal transformation of a piece of information while keeping a privacy-utility tradeoff. More specifically, the contributions of this work are as follows:

- We propose a novel problem statement by identifying three main underlying challenges, namely (1) the assessment on a realisic LLM-based attacker, (2) the sanitisation of the piece of information to limit these inferences and (3) the assessment of the utility of the sanitised output to maintain a threshold;
- We provide a set of design rules to address the above challenges;
- We show experimentally the benefit of following these rules, and conversely the impact of ignoring them, in the context of a hotel review sanitiser.

Paper organization. Section 2 introduces the problem statement. Section 3 introduces the proposed design rules in a case-agnostic fashion. Section 4 describes the application of these rules in the context of sanitising hotel reviews. Section 5 provides an experimental validation and comparison with the state of the art. Section 6 introduces the related work. Lastly, Section 7 concludes the paper and points out future work directions.

2 Problem Statement

This paper tackles the intricate task of text sanitisation through the utilization of LLMs to shield sensitive author attributes (e.g., age, gender, etc.). This Section introduces three underlying difficulties of this task (Sections 2.1, 2.2 and 2.3) and concludes with the problem formulation (Section 2.4).

2.1 Difficulty 1: Using LLMs for Privacy Risk Assessment

Recent studies have shown that off-the-shelf LLMs can infer personal information from texts [16]. However, utilizing pre-trained LLMs in a defensive manner to help individuals assess the risk of inference in their texts presents two main obstacles. Firstly, LLMs can sometimes produce inferences as accurate as a random guess, making them unreliable for privacy risk assessments without a mechanism to evaluate the inference likelihood. Secondly, fine-tuning LLMs has shown effectiveness in various contexts, such as reidentifying personal information from anonymized medical documents [18], highlighting the need for considering high-quality and diverse training datasets to accurately estimate privacy risks.

2.2 Difficulty 2: Assessing Text Utility after Sanitisation

Traditional free-text utility metrics like BLEU [11] and ROUGE [8] may be considered in text sanitisation (see e.g., recent preprint [17]). Such metrics are excellent for evaluating summaries by comparing n-gram co-occurrences between texts. However, they fail to differentiate between the loss of utility due to the modification of words and the destruction of relevant information. For example, consider the following fictitious hotel review and its sanitised version:

> [*Original text*] I went there with my husband Francis for the 3rd anniversary of our youngest child. The staff was delightful and the room clean.

> [*Sanitised text*] Family friendly. The staff was delightful and the room clean.

Although the sanitised version retains all the information essential for evaluating the hotel, an n-gram analysis would result in a low utility score with BLEU, of around 0.27. In addition, consider the identifying excerpt:

> [*Privacy-sensitive excerpt*] I went there with my husband Francis for the 3rd anniversary of our youngest child.

The BLEU value for this Privacy-sensitive excerpt is 0.57, which is surprisingly higher than the score for the sanitized text containing the more descriptive hotel review segment. This discrepancy highlights the challenge of accurately assesing the trade-off between privacy and utility.

2.3 Difficulty 3: Optimising the Overall Sanitisation Process

Designing a text sanitisation process that balances the preservation of content utility with a significant reduction in privacy risks is a multifaceted challenge. Indeed, the inference ability of potential adversaries equipped with fine-tuned LLMs and example datasets must be countered. Existing methods, like masking or direct removal of Personally Identifiable Information (PII), either tend to overly sanitise the text, diminishing its original utility, or retain enough information for sensitives to be inferred. For instance, Microsoft's Azure AI solution [1] can identify text segments that might expose PII and health identifiers (PHI). However, recent studies [16] indicate that simply blacking out these segments is not always effective against inferences made by current LLMs. Yet, these approaches often overlook the utility loss associated with de-identification [3].

2.4 Overall Problem Formulation

The challenge in developing an LLM-based text sanitisation tool is to strike a balance between preserving utility and mitigating privacy risks. To our knowledge, this intricate problem remains unsolved. The objective of this article is to provide a set of design rules to address this goal effectively. This complex task can be distilled into three building blocks:

- **Likelihood measure $\Lambda_{\mathcal{A}}$ for inferences**: assesses the validity of the inferred values for a given set $\mathcal{A}$ of sensitive attributes.
- **Utility measure $\mathcal{U}_{\mathcal{P}}$**: quantifies the utility of a text based on its alignment with a given purpose $\mathcal{P}$ for which the original text was created.
- **Sanitisation process $\mathcal{S}_{\Lambda_{\mathcal{A}},\mathcal{U}_{\mathcal{P}}}$**: transforms the original text into a sanitised text in order to reduce the likelihood of inferences while maintaining utility.

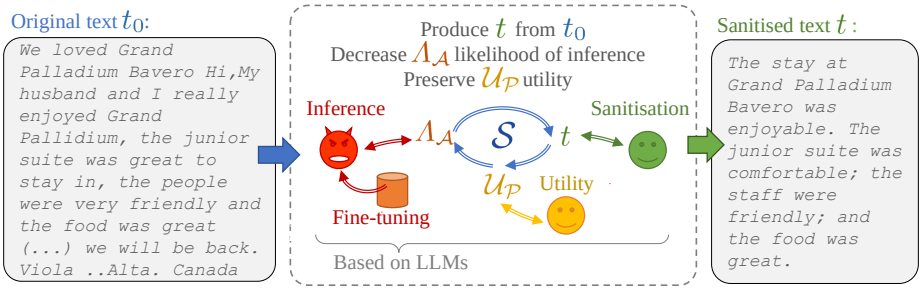


Fig. 1: Main building blocks of a text sanitisation process $\mathcal{S}$ using LLMs

Figure 1 illustrates the overall process and the building blocks for sanitising content using LLMs, from the user text t_0 in input (left) to the sanitised text t in output (right). Our goal is to provide essential guidelines for the design of such LLM-based systems that address the difficulties mentioned above.

3 *reteLLMe* Design rules

In order to address the problem, clear and effective *design rules* are required. This section introduces these design rules, collectively referred to as *reteLLMe*, organised around the three building blocks they address – inference (Section 3.1), utility (Section 3.2) and sanitisation (Section 3.3).

3.1 Inference Design Rules

The ability of LLMs to make inferences can be regarded as an adversary. The issues raised in Section 2.1 require the consideration of a realistic (i.e., sufficiently strong) adversary model suitable for the specific use scenario. An attacker model based on a generic LLM, as considered in recent works [16,17], may underestimate the attacker. This leads to a first practical design rule:

> **Design rule 1: Tailored Adversary LLM.** Avoid using generic attacker models, such as generic LLMs, as this may underestimate accuracy and privacy risks. Instead, employ tailored models such as fine-tuned LLMs.

On the other hand, the capability to infer using LLMs alone is not enough without evaluating the *likelihood* of these inferences. This leads to introduce:

> **Design rule 2: Well-Formed Likelihood Metrics.** The tool must incorporate a well-formed likelihood metrics $\Lambda_{\mathcal{A}}$ to predict the validity of guesses when truth values are unknown.

These design rules imply an attacker model where $\mathcal{RA}$ (Realistic Adversary) has access to a text t authored by a user u and is interested in a specific set of sensitive attributes $\mathcal{A}$. For each sensitive attribute $A{\in}\mathcal{A}$ (e.g., Age), D_A represents its domain, and $a_u{\in}D_A$ denotes its true value for u (e.g., the actual age of u). Using LLMs, $\mathcal{RA}$ produces a_t, the value it inferred from t for each a_u. To represent realistic threats, the adversary is assumed to have access to:

- a dataset D of pre-existing texts written by a set of users U, not including u, for which real values of sensitive attributes $a_{u' \in U}$ are known;
- a likelihood metric $\Lambda_{\mathcal{A}}$ which evaluates the accuracy of guesses about sensitive attributes (i.e., the probability that the inferred value a_t matches the true value a_u).

Designing $\Lambda_{\mathcal{A}}$ is challenging. The likelihood metric estimates the accuracy of each guess made by the attacker. Ideally, a guess represents the probability of its correctness. However, since the truth values of targeted users during the attack are unknown, this probability cannot be analytically computed.

A likelihood metric $\Lambda_{\mathcal{A}}$ is considered well-formed and satisfies design rule 2, if and only if it satisfies the following property: for u the author of text t, $\forall A \in \mathcal{A}$,

$\forall \epsilon \in [0,1], \quad \epsilon \mapsto \frac{|\{t \in D : \Lambda_{\mathcal{A}}(a_t) > \epsilon \wedge a_t = a_u\}|}{|\{t \in D : \Lambda_{\mathcal{A}}(a_t) > \epsilon\}|}$ is a monotonically increasing function. It ensures that $\Lambda_{\mathcal{A}}$ can predict whether the probability of a guess being correct is above or below a given threshold. This capability allows an attacker to know whether the inference is plausible and it allows a sanitising tool to alert the user that their text is at risk.

3.2 Utility Assessment Design Rules

We advocate for the adoption of purpose-centric utility metrics alongside inference metrics. The objective is to identify and prioritize information aligned with the intended *purpose* of texts. This leads to the following rule:

Design rule 3: Purpose-Centric Utility. The integration of purpose-centric utility metrics $\mathcal{U}_{\mathcal{P}}$, defined independently of privacy considerations and tailored to the specific purpose of the original text, is essential for maintaining the practical value of LLM-based sanitised outputs.

This guideline entails defining a purpose as a set $\mathcal{P}$ of purpose-related attributes, where each attribute $P \in \mathcal{P}$ represents a category of relevant information regarding the initial purpose for which the text was produced.When a text t produced by sanitising t_0 transmits information about an attribute linked to a purpose $\mathcal{P}$, other users can deduce a value from t. $\mathcal{U}_{\mathcal{P}}$ should evaluate the ability of t to convey *the same* information relevant to the purpose as t_0.

3.3 Sanitisation design rules

The process of sanitisation relies on LLMs to transform a text t_0 into a sanitised text t, aiming to reduce privacy risks while maintaining utility. We do not prescribe specific guidelines for the sanitisation process, whether it should be iterative, interactive, or otherwise.

The effectiveness of a sanitisation technique hence hinges on the independence between the purpose-centric attributes $\mathcal{P}$ and the sensitive attributes $\mathcal{A}$, leading to a fourth rule:

Design rule 4: Privacy-Utility Independance. Sanitisation techniques must aim to decrease inference likelihood while retaining useful information. The efficiency of the sanitisation process is constrained by the degree of independence between privacy and utility metrics. In case where independence is lacking, residual privacy risks must be carefully evaluated and addressed.

The theoretical feasibility of a perfect sanitiser, which would fully preserve utility while nullifying privacy threats, relies on the independence of relevant information categories between sensitive attributes and purpose-centric ones. However, in practical scenarios where independence is not assured, different trade-offs

need exploration, and residual privacy risks must be considered. This underscores the importance of robust privacy risk assessment, as highlighted in Section 3.1.

4 Application of *reteLLMe* design rules

Demonstrating the suitability of the *reteLLMe* design rules requires instantiating them in a practical scenario. This section shows the instantiation of a sanitiser based on an LLM for sanitising hotel reviews. The scenario and dataset are described in Section 4.1. Then the design rules for inference, utility assesment and sanitisation are implemented in Section 4.2.

4.1 *"Hotel reviews sanitiser"*: scenario and dataset description

We consider a practical application scenario where a sanitisation tool based on $ChatGPT3.5$ is at stake – users enter their hotel review text and the tool rewrites the text to improve privacy while preserving the review utility. This tool could be integrated into commercial platforms such as Booking.com or Airbnb.

To validate our design rules, we use the PAN[6] dataset [13], which provides 4.160 hotel reviews written in English (see an example on left part of Figure 1). Each one has truth values of two attributes of its author – *gender* (male or female) and *age* ($[18, 24]$, $[25, 34]$, $[35, 49]$, $[50, 64]$ or $[65, xx]$).

4.2 Implementation and compliance to design rules

The inference values a_t for age (A) and g_t for gender (G) for a text t are produced using a specific prompt. The inference process involves fine-tuning $ChatGPT3.5$ using a random subset of the PAN hotel reviews dataset. Concerning likelihood values, they are also produced using specific prompts. All prompts are shared in our online repository as stated below.

In what comes to utility, we define a set of purpose-related attributes $\mathcal{P}$ that typically summarise hotel reviews, including general sentiment, specific problem noted, cleanliness, room quality and service standards. Each attribute in $\mathcal{P}$ is categorized into positive (*good*), negative (*bad*), or neutral/missing ($\perp$) values, reflecting its impact on the overall assessment of hotel performance. For each $P \in \mathcal{P}$, we define a binary utility, which is 1 if $ChatGPT3.5$ provides the same answer for t and t_0 (the latter being a non-null value), and 0 otherwise. The overall utility $\mathcal{U}_{\mathcal{P}}$ is computed as the average of these utilities. Concerning sanitisation, $ChatGPT3.5$ is instructed to eliminate textual elements that could reveal the reviewer's age or gender. Subsequently, the anonymised text is rewritten in a neutral tone to mitigate unintended biases.

[6] PAN is an annual competition [12] that provides datasets for different tasks, including author profiling. We are using the 2014 dataset which is the most comprehensive provided for our use. PAN also provides the accuracy achieved by the winners, which will be used for comparison.

All these decisions are in line with design rules: we apply fine-tuning (Design rule 1); the likelihood of each inference is also self-evaluated by $ChatGPT3.5$, and we show experimentally in the next section that the resulting likelihood metric is well-formed (Design rule 2); we score texts depending on their information across purpose-related attributes (Design rule 3) and we balance privacy and utility in sanitized texts (Design rule 4).

5 Assessment of *reteLLMe* design rules

This section presents the experimental results on the proposed $reteLLMe$ design rules. First, the methodology and experimental settings are discussed in Section 5.1. Afterwards, we use the same order as in the previous sections – Section 5.2 focuses on the inference process, Section 5.3 on utility computation, Section 5.4 on the sanitisation effectiveness. Lastly, Section 5.5 discusses the results.

5.1 Experimental settings

Since the attacker model involves fine-tuning which may lead to variations depending on the training dataset, we randomly partition the dataset in four. Each partition is randomly split into two categories, training (H_{840}, approximately 80% of texts) and tests (the remaining 20%). Experiments are run four times on each dataset. Reported results are the average of all experiments.

To generate our prompts, we used well-known techniques (e.g., "Let's play a game...") to force ChatGPT[7] to perform undesired actions [7,16]. To fine-tune ChatGPT, we used OpenAI's dedicated API [10]. To foster further research, our experimental scripts and prompts are publicly released[8].

5.2 Validating reteLLMe measure for inference and likelihood

The goal of this section is to validate the inference module by (i) evaluating the impact of fine-tuning ChatGPT and confirming the strength of the realistic attacker (to assess Design rule 1) and (ii) validating the proposed likelihood metric (to assess Design rule 2).

Attacker definitions. In accordance with Design rule 1, our experiment considers a (realistic) strong attacker relying on a fine-tuned version of ChatGPT as described above. For comparison, this *strong attacker* is compared to a *weak attacker* that behaves similarly but relies on an off-the-shelf ChatGPT.

[7] We used $ChatGPT3.5$.
[8] GitHub repository to be added after acceptance

Fine-tuning - Design rule 1. Figure 2 shows the accuracy of the strong attacker and its weak version with respect to a random guess and the best scores in each category of the PAN competition. The random guess ("baseline") has a score of 0.5 for gender as there are only two options (male/female), and 0.2 for the age as there are five ranges (see Section 4.1). Thus, the *total* baseline accuracy is 0.1 as the product of the two.

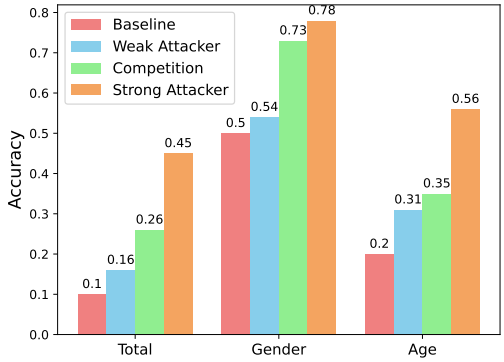
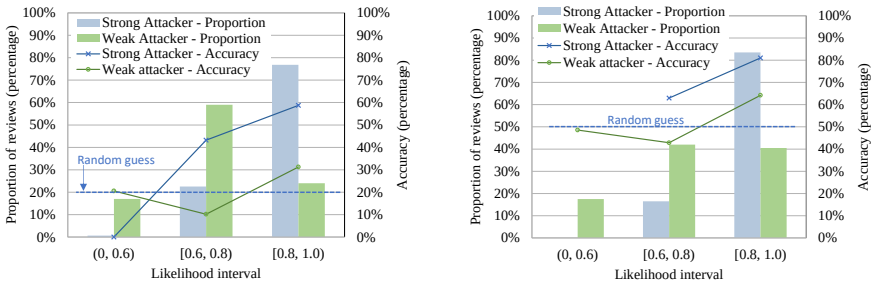


Fig. 2: Accuracy of the weak and strong *reteLLMe*-compliant attackers

Figure 2 shows a significant improvement in the total accuracy from 0.16 to 0.45 with fine-tuning, surpassing both the baseline and the competition results. Furthermore, the comparative analysis between age and gender reveals that the gender category benefits more from fine-tuning than the age category. This could be explained by the differences in the complexity of inferring age, which involves multiple categories, versus gender, which is classified as female or male.

Well-formed likelihood - Design rule 2. Figure 3 shows the number of inferences and their average accuracy as a function of their likelihood range. Each attacker provides its own likelihood, so the inferences are partitioned twice according to both. For each likelihood range, the bars show the proportion of inferences (percentage of reviews, left axis) and the curves show their average accuracy (right axis). Note that when no review lies in a likelihood interval, the value of the curve representing accuracy cannot be computed (as in the case of (0,0.6) for gender with the strong attacker).

Accuracy of the strong attacker for age and gender shows a correlation with likelihood (Pearson value of 0.99 for age and 0.96 for gender), as opposed to the weak attacker (age: 0.33, gender: 0.56). In fact, for the strong attacker, the accuracy increases monotonically for increasing likelihood ranges, showing that the likelihood metric is in this setting a well-formed metric (in the sense of Section 3.1) that satisfies *Design rule 2*.



(a) Age (b) Gender

Fig. 3: Accuracy of inference based on likelihood (original texts)

Underestimation of risks - Design rule 1. Beyond the well-formedness, Figure 3 illustrates the distribution of text accross different likelihood levels. A text is at risk when it falls within a likelihood range where the average accuracy significantly exceeds random guessing. Here, in reality, all texts are at risk. The (realistic) strong attacker, adhering to $reteLLMe$ design rules, achieves a likelihood greater than 0.6 for each text, with an average accuracy well above random guessing, even within the $[0.6, 0.8)$ likelihood interval. On the contrary, the weak attacker fails to identify texts with likelihood below 0.8. This discrepancy leads to a significant underestimation of risks: 76% of texts (for age) and 59.5% (for gender) with high inference risks would not be flagged as such.

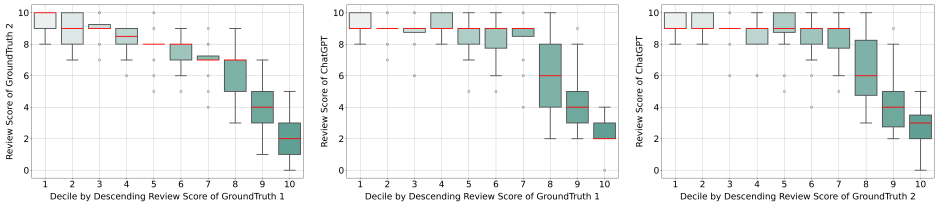
5.3 Validating $reteLLMe$ measure for utility

This section validates the purpose-centric $reteLLMe$ utility measure presented above (see Section 4.2). It first compares the responses obtained using this measure with those provided by humans. It then shows how BLEU and ROUGE behave and concludes on the importance of *Design rule 3*.

Automated purpose-related utility - Comparison with humans. We assign to each review a score out of 10. The score is calculated as the sum of each of the five purpose-related attributes, by assigning 2, 1, and 0 points to each "good", "neutral" and "bad" value respectively. This score is therefore not a measure of the utility of the review (a negative review can be very useful), but simply a way of ranking the reviews from the most positive to the most negative.

Figure 4 shows the distribution of scores for each of the humans and ChatGPT responses. Each box represents 10% of reviews, from the most positive to the most negative. These three curves show a strong similarity between the humans themselves and between the humans and ChatGPT. This leads to Pearson correlations of 0.8 and 0.82 between each human and ChatGPT. However, it should be noted that the variations in ChatGPT are less uniform than those

between humans. Therefore, although imperfect, ChatGPT is a good source of utility in the absence of ground truth.



(a) Ground truth 1 vs Ground truth 2 (b) Ground truth 1 vs ChatGPT (c) Ground truth 2 vs ChatGPT

Fig. 4: Human and ChatGPT-based evaluation of purpose-related attributes

BLEU/ROUGE against purpose-related utility - Design rule 3. To analyse the suitability of BLEU and ROUGE metrics, we examine their alignment with purpose-related utility metric by applying these measures to original hotel reviews (t_0) and sanitised versions (t) and comparing obtained scores with utility preservation according to our proposal.

Figure 5 presents the distribution of BLEU and ROUGE scores, categorized into quintiles with increasing utility preservation. Utility preservation for each sanitized text t compared to original text t_0 is determined by analysing $ChatGPT3.5$'s responses to a purpose-related questionnaire using both texts. This involves calculating binary utility $\mathcal{U}_P$ for purpose-related attribute based on $ChatGPT3.5$'s responses (which is 1 is same answer is provided, and 1 otherwise) and averaging these values for the five attributes to derive utility preservation (i.e., 1 means fully preserved with same answers to the five purpose-related questions, 0 means answers are all different).

Our results show a significant decorrelation between BLEU/ROUGE scores and utility preservation. Hence, they inadequately assess the purposeful information conveyed in texts, highlighting the negative repercussion of disregarding *Design rule 3*.

5.4 Sanitisation effectiveness

We compare our method to the two settings of the anonymiser proposed by Azure, Azure (All entities) and Azure (Three entities) [1]. There are three issues to consider – whereas after sanitisation the inference likelihood is effectively reduced, the inference accuracy is also decreased and the utility is preserved, as needed to satisfy *Design rule 4*.

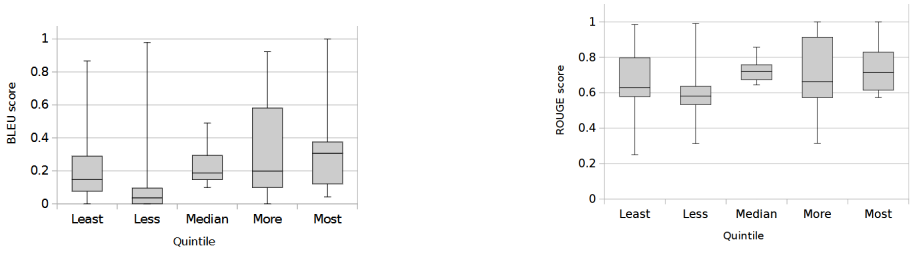
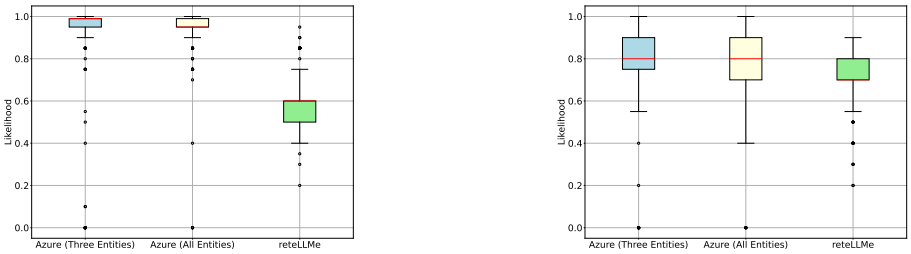


Fig. 5: BLEU/ROUGE scores of sanitised texts ordered by purpose-centric utility

Decreasing likelihood. Figure 6 shows the distribution of the inference likelihood for both attributes in the three considered methods. Intuitively, lower likelihood values are preferred from a privacy perspective.



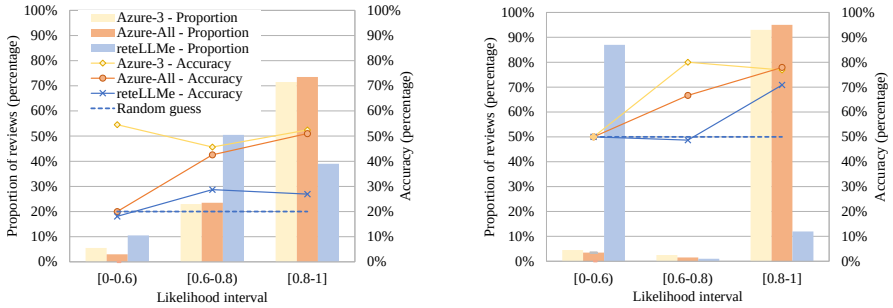
(a) Gender (b) Age

Fig. 6: Inference likelihood after sanitisation

Figure 6a illustrates the inability of Azure to sanitise gender-related data in both settings, resulting in median likelihood scores of 0.99 and 0.95. Unsurprisingly, the setting removing less data, "Three entities", exhibits the highest likelihood score. On the contrary, *reteLLMe* is significantly more effective to protect this attribute, leading to a median likelihood of 0.6. A similar situation happens for the attribute age (Fig. 6b). Thus, our module outperforms both settings of Azure for the protection of both age and gender attributes.

Decreasing accuracy. Figure 7 shows the distribution of accuracy values for the three methods. Recall that random guess thresholds are different for age (which is 0.2) and gender (being 0.5). Intuitively, lower likelihood scores lead to smaller accuracy values. Interestingly, *reteLLMe* exhibits good behavior for the highest level of likelihood. Thus, an average accuracy of 0.27 and 0.71 is reached for the age and gender, respectively. For lower likelihood ranges, *reteLLMe* accuracy is closer to a random guess.

As a matter of fact, the amount of texts that remain at risk after sanitisation is largely different. Remarkably, for both attributes and both Azure variants, more than 90% of reviews remain at risk, i.e. belongs to likelihood intervals with an average accuracy significantly more than random guessing (up to 52% average accuracy for age and 78% for gender). When *reteLLMe* is applied, only 11% of reviews are at risk with regard to gender.



(a) Age (b) Gender (same legend)

Fig. 7: Accuracy of inference based on likelihood (after sanitisation)

Utility preservation. Beyond reducing the likelihood of inferences, it is also necessary to ensure that the utility is preserved. Figure 8 shows the distribution of utility preservation across the three methods at stake. The three methods lead to a substantial utility preservation, as the highest amount of records count on the biggest utility preservation figures. Indeed, *reteLLMe*, Azure 3 entities, and Azure all entities preserve between 80% and 100% utility of 71.6%, 70,5%, and 65.6% of texts respectively. Overall and as expected, the "3 entities" setting outperforms the "all entities" setting with regard to utility preservation.

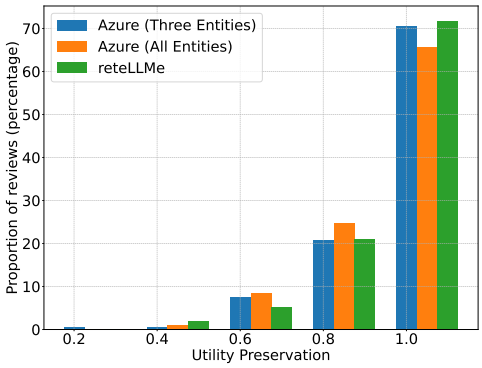


Fig. 8: Distribution of utility preservation

5.5 Discussion

Our experimental results are merely limited to the use-case of hotel reviews with a limited test dataset (4 experiments with 200 texts per experiment). They are however valid to confirm the proposed guidelines.

First, they confirm that LLMs such as ChatGPT can be used as privacy-enhancing tools in an effective manner that outperforms industrial state of the art anonymisers in term of both risk minimization and utility preservation. More importantly, they demonstrate the importance of our design rules.

Indeed, they illustrate the effect of considering a strong adversary by showing how fine-tuning impacts ChatGPT's inference ability. They demonstrate how generic adversaries may severely underestimate privacy risks.

As opposed to the state-of-the-art, they shown that purpose-centric utility metrics are a differentiating factor. Specifically, it is shown that generic metrics such as ROUGE or BLEU may be wholly decorrelated to the amount of purposeful information present in a text.

Our results, however, are limited in that Design rule 4 asked for a sanitisation procedure that decreases the inference likelihood while preserving utility. Nevertheless, our input to ChatGPT does not include the inference likelihood. Our results confirm that even without that input, the inference likelihood and accuracy are severely decreased while the utility is preserved. We argue that this phenomenon cannot be extrapolated to any use case, as it may be due to the very nature of hotel reviews. Nevertheless, our assessment is comprehensive enough to confirm that even the most simplified sanitiser observing this design rule is effective enough.

6 Related Work

The use of LLMs as privacy-enhancing technologies has already attracted some research attention. Indeed, an increasing amount of papers have been produced in the last years. Interested readers may refer to a systematic literature review by Sousa *et al.* [15]. In a nutshell, LLMs seem to be a suitable technology considering the challenges posed by text anonymization, due to the unbounded nature of information related to individuals [9].

[16] was the first study to highlight the critical privacy concerns posed by LLMs beyond the commonly discussed issues related to data memorization. They show that LLMs are able to identify personal data at an unprecedented scale and emphasizes the need for new anonymization techniques to counteract such evolving threats. While the authors effectively illustrate the problem the of inferring personal data from text, they do not propose any solution. Our research builds upon their findings by not only considering these issues but also proposing new guidelines to mitigate the risk of such inferences. Moreover, contrary to our guidelines, they count on a generic attacker – their LLM is not fine-tuned.

A follow-up work by the same authors [17] marks a significant advancement in this domain. They propose an evaluation framework that leverages the capabilities of LLMs for text anonymization. It employs a multiple round process

where a LLM adversary analyzes the text for private attribute inference, followed by an anonymizing LLM that modifies the text to obscure identifiable information. Furthermore, this framework introduces a binary "certainty" scoring system discriminating inferences depending on whether they rely on statistical bias or directly identifiable information within texts. We consider that this "certainty" score serves as a metric of inference likelihood. However, [17] does not validate its relation to accuracy. Since statistical analysis may provide accurate guesses, it is not immediate that certainty is a good predictor of accuracy.

Beyond the inference likelihood, Staab *et al.* consider utility metrics such as BLEU and ROUGE. Complementarily, they use a "judge" prompt that assesses anonymized texts across three dimensions – readability, meaning, and hallucination. Interestingly, the "meaning" dimension assesses semantic proximity that could be purpose-centric. Since it takes into account *all* information, we argue that it exhibits the same limitations as those discussed in Section 2.2 and can never be independent from privacy considerations. This is supported by the observation that BLEU, ROUGE, and judge exhibit the same trends [17].

Another relevant study is [4]. The authors introduce the concept of self disclosure abstraction that allows paraphrasing personal disclosures into more general terms without losing their communicative value, reducing privacy risks while preserving the overall utility (e.g., "Im 16F" to "I'm a teenage girl".) Their methodology involves a fine-tuning strategy to identify instances of self-disclosures which confirms the importance of considering a realistic attacker model as highlighted in our guidelines. However, their approach does not work properly for sensitive attributes which are not directly mentioned in the text, as [16] already proved.

[18] addresses the issue of the de-identification of clinical reports to facilitate data access for research purposes while ensuring patient privacy using the CamemBERT model, a BERT variant specially crafted for French texts. This approach aligns partially with our proposed guidelines. They count on a well-formed inference likelihood metric. Moreover, their attacker model is concrete enough due to fine-tuning. However, it does not include the notion of utility.

Our work distinguishes itself from recent research by proposing a number of guidelines that have been partially overlooked by previous efforts, as discussed above. To the best of our knowledge, this is the first effort in this direction. Our work has also illustrated which is the impact of not following these guidelines.

7 Conclusion

LLMs have already been shown to be effective to both anonymise and de-anonymise texts. This dual nature gives them an unprecedented ability to be used as privacy-enhancing technology. However, previous attempts have failed to propose such an usage considering the common pitfalls in text utility, inference assessment and sanitisation effectiveness. In this vein, this work has proposed *reteLLMe*, a collection of design rules in this regard. Our assessment in the context of protecting hotel reviews has not only shown the convenience of the proposed rules, but also the negative consequences of disregarding them.

Future work will focus on exploring the suitability of these rules in other contexts. In this vein, the design of well-formed and generalizable purpose-centric utility metrics is envisioned as a critical issue. On the other hand, exploring the impact of LLM-based threats such as privacy leakages [6] in this privacy-enhancing usage is another interesting direction.

Acknowledgement

This work was supported by the French grant iPoP PEPR (ANR-22-PECY-0002). Jose Maria de Fuentes has been partially supported by the Spanish National Cybersecurity Institute (INCIBE) grant APAMciber within the framework of the PRTR funds, financed by the European Union (Next Generation). Jose Maria de Fuentes has also received support from UC3M's Requalification programme, funded by the Spanish Min. de Ciencia, Innovacion y Universidades with EU recovery funds (Convocatoria de la UC3M de Ayudas para la recualificación del sistema universitario español para 2021-2023, de 1 de julio de 2021).

References

1. Azure: What is Azure AI Language? https://learn.microsoft.com/en-us/azure/ai-services/language-service/overview (July 2023), [Online; accessed 01-April-2024]
2. Bai, J., Men, R., Yang, H., Ren, X., Dang, K., Zhang, Y., Zhou, X., Wang, P., Tan, S., Yang, A., et al.: Ofasys: A multi-modal multi-task learning system for building generalist models. arXiv preprint arXiv:2212.04408 (2022)
3. Berthelier, G., Boutet, A., Richard, A.: Toward training nlp models to take into account privacy leakages. In: 2023 IEEE International Conference on Big Data (BigData). pp. 4854–4862. IEEE (2023)
4. Dou, Y., Krsek, I., Naous, T., Kabra, A., Das, S., Ritter, A., Xu, W.: Reducing privacy risks in online self-disclosures with language models. arXiv preprint arXiv:2311.09538 (2023)
5. Kandpal, N., Pillutla, K., Oprea, A., Kairouz, P., Choquette-Choo, C., Xu, Z.: User inference attacks on llms. In: Socially Responsible Language Modelling Research (2023)
6. Kim, S., Yun, S., Lee, H., Gubri, M., Yoon, S., Oh, S.J.: Propile: Probing privacy leakage in large language models. Advances in Neural Information Processing Systems **36** (2024)
7. Li, H., Guo, D., Fan, W., Xu, M., Huang, J., Meng, F., Song, Y.: Multi-step jailbreaking privacy attacks on ChatGPT. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 4138–4153. Association for Computational Linguistics (2023)
8. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
9. Lison, P., Pilán, I., Sánchez, D., Batet, M., Øvrelid, L.: Anonymisation models for text data: State of the art, challenges and future directions. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 4188–4203 (2021)

10. OpenAI: Openai documentation, https://platform.openai.com/docs/guides
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
12. Pardo, F.M.R., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at PAN 2013. In: Forner, P., Navigli, R., Tufis, D., Ferro, N. (eds.) Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013. CEUR Workshop Proceedings, vol. 1179. CEUR-WS.org (2013), https://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-RangelEt2013.pdf
13. Rangel Pardo, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd author profiling task at pan 2014. CEUR Workshop Proceedings **1180**, 898–927 (2014)
14. Song, Y., Zhang, J., Tian, Z., Yang, Y., Huang, M., Li, D.: Llm-based privacy data augmentation guided by knowledge distillation with a distribution tutor for medical text classification. arXiv preprint arXiv:2402.16515 (2024)
15. Sousa, S., Kern, R.: How to keep text private? a systematic review of deep learning methods for privacy-preserving natural language processing. Artificial Intelligence Review **56**(2), 1427–1492 (2023)
16. Staab, R., Vero, M., Balunović, M., Vechev, M.: Beyond memorization: Violating privacy via inference with large language models. arXiv preprint arXiv:2310.07298 (2023)
17. Staab, R., Vero, M., Balunović, M., Vechev, M.: Large language models are advanced anonymizers. arXiv preprint arXiv:2402.13846 (2024)
18. Tannier, X., Wajsbürt, P., Calliger, A., Dura, B., Mouchet, A., Hilka, M., Bey, R.: Development and validation of a natural language processing algorithm to pseudonymize documents in the context of a clinical data warehouse. Methods of Information in Medicine (2024)
19. Zhang, X., Xu, H., Ba, Z., Wang, Z., Hong, Y., Liu, J., Qin, Z., Ren, K.: Privacyasst: Safeguarding user privacy in tool-using large language model agents. IEEE Transactions on Dependable and Secure Computing (2024)