Sanitization or Deception? Rethinking Privacy Protection in Large Language Models

Bipin Paudel
Kansas State University
Manhattan, Kansas, USA
bipinp@ksu.edu

Bishwas Mandal
Kansas State University
Manhattan, Kansas, USA
bishmdl76@ksu.edu

George Amariucai
Kansas State University
Manhattan, Kansas, USA
amariucai@ksu.edu

Shuangqing Wei
Louisiana State University
Baton Rouge, Louisiana, USA
swei@lsu.edu

Abstract

Large language models have shown considerable abilities across many tasks, but their capacity to detect sensitive user information from text raises significant privacy concerns. While recent approaches have explored sanitizing text to hide private features, a deeper challenge remains: distinguishing true privacy preservation from deceptive transformations. In this paper, we investigate whether LLM-based sanitization reduces private feature leakage without misleading an adversary into confidently predicting incorrect labels. Using LLM as both sanitizer and adversary, we measure leakage using two entropy-based metrics: Empirical Average Objective Leakage (E-AOL) and Empirical Average Confidence Boost (E-ACB). These allow us to quantify not only how accurate adversarial predictions are, but also how confident they remain post-sanitization. We posit that deception, while reducing adversarial accuracy, will also increase confidence in incorrect inferences, and hence reduced accuracy alone should not be interpreted as true privacy. We show that while current LLMs can hide private features, their transformations sometimes cause deception. Finally, we evaluate the semantic utility of sanitized outputs using sentence embeddings, LLM-based similarity judgments, and standard metrics like BLEU and ROUGE. Our findings emphasize the importance of explicitly distinguishing between privacy and deception in LLM-based sanitization and provide a framework for evaluating this distinction under realistic adversarial conditions.

Keywords

textual privacy, privacy vs deception, min-entropy metrics

1 Introduction

Recently, there has been a rapid expansion in AI-driven sectors, fueled by advancements in architectural designs, increased model parameter sizes, the widespread availability of open-source models, access to vast datasets, and other contributing factors. These developments are revolutionizing industries such as finance, education, entertainment, and e-commerce [33].

In particular, the emergence of transformer-based models [42] has led to the creation of large language models (LLMs) with immense parameter counts, such as GPT-3, which boasts 176 billion parameters [5], and its successor GPT-4 [1], which incorporates multimodal capabilities, including image interpretation. These LLMs exhibit a broad range of abilities to comprehend human-level information and execute tasks across various domains through zero-shot and few-shot learning paradigms [5].

While many of us are enthusiastic about how artificial intelligence is enhancing our lives, the rapid advancements in AI also raise significant privacy concerns. A primary issue is that large language models (LLMs) are trained on vast amounts of online data, including potentially sensitive information, which these models can inadvertently memorize [15, 17, 29]. This memorization could unintentionally leak sensitive user data such as social security numbers, addresses, phone numbers, emails, and more. Additionally, a new dimension of privacy leakage is explored by the authors of [38], who demonstrate that LLMs are capable of inferring private attributes like income, gender, relationship status, occupation, and location based on the information users share online, raising significant concerns about inference-based privacy leakage.

To mitigate these risks, recent work has explored techniques for sanitizing user-generated content, aiming to preserve the utility of text while minimizing the ability of adversaries to infer private information [26, 39]. While early approaches targeted data privacy, such as preventing memorization [17, 19, 20, 22], more recent efforts seek to obscure attributes that can be inferred from natural language. One emerging line of work proposes prompt-based frameworks to systematically transform user inputs.

Yet, an important and underexplored challenge remains: distinguishing between privacy-preserving transformations and those that introduce deception. Consider the following motivating example. Suppose a male user wants to hide their gender. After sanitization, the model replaces a male emoji with a female emoji, leading an adversary to infer "female" with high confidence. While this may prevent correct inference, it does not just hide the original feature, it actively misleads the adversary. This raises a fundamental question:

> *Can LLMs effectively sanitize private information without misleading the adversary into confidently predicting the wrong label?*

Proceedings on Privacy Enhancing Technologies 2026(1), 154–174

https://doi.org/10.56553/popets-2026-0009

The distinction between privacy and deception is critical in contexts where downstream decisions (e.g., advertising, employment screening) rely on inferred features. A system that simply "flips" labels rather than reducing inference confidence may give the illusion of privacy while introducing new risks. This distinction is especially important under asymmetric knowledge, where the adversary does not know the exact sanitization mechanism, making it difficult to detect whether the text has been modified to hide or to mislead.

In this paper, we build on the entropy-based framework introduced in [34–36], which quantifies privacy leakage through metrics such as Average Objective Leakage (AOL), Average Confidence Boost (ACB), and Average Subjective Leakage (ASL). These metrics are well-suited for capturing privacy risk under realistic, partially informed adversaries. Using these tools, we empirically study how LLMs behave under different prompting strategies, whether asked to preserve privacy or explicitly mislead, and how that behavior impacts adversarial accuracy and confidence.

To enable this analysis, we adopt and enhance a single-step, self-verifying prompt-based sanitization technique. While not the core contribution of this paper, this method improves upon prior iterative frameworks (e.g., [39]) by reducing model interaction and computational overhead. It allows us to isolate and analyze how LLMs internalize instructions and whether the resulting transformations truly preserve privacy or introduce deception.

The main contributions of this paper are as follows:

- **Privacy vs. Deception Distinction:** We formalize the boundary between privacy-preserving and deceptive transformations in text sanitization, showing that reductions in adversarial accuracy do not necessarily imply privacy unless adversarial confidence also decreases.
- **Empirical Evaluation Framework:** We design an experiment where LLMs act as both sanitizer and adversary, and use entropy-based metrics (E-AOL, E-ACB) to measure inference risk before and after sanitization or deception.
- **Quantitative and Statistical Analysis:** We quantify the difference in leakages based on confidence intervals and perform statistical testing to assess when sanitization reduces leakage, and increase confidence after deception.
- **Comparative Utility Assessment:** To evaluate the utility of the sanitization mechanism, we compute sentence-level embeddings to measure semantic similarity between the original and sanitized text. Additionally, we employ standard natural language processing metrics such as ROUGE and BLEU to assess content preservation. To further capture contextual similarity, we incorporate a language model based utility measure, where a large language model is asked to evaluate the similarity between the original and sanitized sentences in terms of both meaning and readability of the sanitized text.

2 Background and Related Work

Foundations of Privacy in Machine Learning Privacy in machine learning is commonly categorized into data privacy and inference privacy [40, 41]. While data privacy focuses on securing raw inputs, inference privacy aims to prevent the unintended derivation of private information. It is now well understood that protecting the input data alone is not sufficient to guarantee privacy at the inference level.

The vast datasets used to train deep learning models, which comprise billions of parameters, make the issue of privacy of critical importance. Since these models are widely employed across various sectors, they are susceptible to numerous privacy attacks. Membership inference attacks attempt to determine whether a specific example was included in the training set [18, 32], whereas property inference attacks focus on uncovering attributes of the training distribution [2, 14, 50]. Model extraction attacks aim to replicate or reconstruct an entire model by querying the target model [11, 21], while model inversion attacks attempt to reconstruct training data points based on the model's outputs [13, 51].

Privacy Breach Scenarios With the rise of modern large language models, training data extraction attacks have become increasingly prominent in the literature. These attacks are similar to model inversion attacks but aim to reconstruct verbatim training examples rather than just representative samples [6]. Several studies [15, 17, 29, 48] demonstrate that due to memorization in large language models, it is possible to extract exact sentences from the training data. Such extracted information may reveal private information about users, including email addresses, phone numbers, addresses, or even Social Security numbers. The authors of [17] categorize training data extraction attacks into two types: targeted and untargeted attacks. Targeted attacks focus on extracting specific segments of text, such as emails, phone numbers, or URLs. In contrast, untargeted attacks aim to recover entire training examples, leading to the theft of sensitive and valuable private data.

Inference Privacy and Adversarial Mitigation While there are several approaches in mitigating certain aspects of data privacy [17, 19, 20, 22], the challenge of inference privacy remains [40, 41]. Existing research points out that the advanced capabilities of machine learning models pose significant privacy concerns. For example, the authors of [38] illustrate that these concerns extend beyond simple memorization, showing how advancements in large language models (LLMs) can be exploited to infer private user information. The study demonstrates that the statistical knowledge embedded in LLMs is not limited to textual data but also poses privacy risks for tabular datasets. To address this, [26] proposes a method for sanitizing tabular datasets using LLMs to obscure private features, such as income status, while preserving the utility of the data.

Various adversarial optimization techniques have been developed to address inference privacy by distorting datasets to maintain privacy while preserving utility [12, 16, 25, 31, 37, 45, 47]. These approaches primarily involve introducing noise into the generator network, modifying the latent representations of inputs, or applying special loss functions to obscure private features. These methods are predominantly used with tabular datasets, images, or text representations. Notably, [25] highlights the significance of additional post-processing tailored to the specific characteristics of the dataset.

Privatizing textual data using transformer-based models is a growing area of research, with several approaches aiming to mitigate the risk of de-identifying individuals through their written

content. Prior work has addressed the challenge of authorship de-anonymization by targeting stylistic cues that are unique to individuals. For example, the authors in [27] explore the limits of privacy in pretrained language models by analyzing how perturbations, such as noise added at the final layer, affect the ability of an adversary to infer sensitive attributes. In contrast, the authors in [3], introduce a reinforcement learning based framework that fine-tunes language models to rewrite text in a way that obfuscates author identity while preserving fluency and meaning.

Beyond authorship obfuscation, other works have explored protecting sensitive attributes embedded in user prompts sent to large language models (LLMs). For example, the authors in [9] propose applying differential privacy-based text sanitization prior to sending prompts to untrusted LLM providers, coupled with a pre-sanitization utility assessment to ensure that transformations do not excessively degrade task-relevant content. Similarly, [7] presents a middleware framework that uses a trusted small language model to predict the utility of sanitized prompts before transmitting them to costly or untrusted LLMs, thereby reducing resource waste while preserving privacy.

Complementary to these privacy-preserving approaches, the authors in [43] audit commercial generative AI assistants to investigate how they perform tracking, profiling, and personalization during user interactions. Their analysis reveals that AI assistants may accumulate sensitive behavioral and demographic data over time, enabling fine-grained profiling that could be exploited for targeted advertising or other purposes. Such findings highlight the broader risks associated with user-LLM interactions, reinforcing the importance of mechanisms that limit unintended data leakage.

While these works focus on deanonymizing the author, our privacy objective differs. We are specifically interested in hiding users' private features, like their income, gender, relationship, and age-related information, from the textual information they provide to the internet. These attributes are considered as Personally Identifiable Information(PII) by the EU's General Data Protection Regulation (GDPR) and require such data to be protected. Similarly, there are US-centric regulations such as California's CCPA, which require such personal identifiable information (PII) to be reasonably inferred by either direct or indirect means.

This text-based privacy issue is also explored in [39], which investigates prompt-based sanitization methods for obscuring private features and evaluates the impact on adversarial inference accuracy. The study suggests that a reduction in adversarial accuracy is a key indicator of effective privacy protection. However, we argue that this metric alone may be insufficient. If a model introduces misleading cues that deliberately steer the adversary toward an incorrect label, the transformation may no longer be deemed as private but rather deceptive, which has risks of its own.

Quantifying Privacy and Deception As large language models are increasingly used for sanitization tasks, whether on textual data [39] or tabular datasets [26], it becomes essential to distinguish between genuine privacy preservation and unintended deception. While aiming to obscure private features, a sanitization mechanism should not alter the text in ways that deliberately mislead an adversary into confidently predicting an incorrect label. Such behavior may reduce adversarial accuracy, but not necessarily in a privacy-preserving manner.

To address this, the works in [34–36] explore the relationship between privacy and deception using both Bayesian and non-Bayesian frameworks. These studies introduce several metrics based on min-entropy leakage, such as Average Confidence Boost (ACB), Average Objective Leakage (AOL), and Average Semantic Leakage (ASL), to quantify privacy risks based on the adversary's confidence after observing partial system outputs. One limitation of many existing studies is the assumption that adversaries have complete knowledge of the privacy mechanism, which is rarely the case in real-world applications. These metrics instead help assess adversarial inference under partial information, providing a more realistic perspective.

In this paper, we extend the sanitization framework introduced by [39] on text-based data and incorporate min-entropy based key metrics from [36] to develop a more efficient and effective sanitization strategy. Our approach aims to distinguish between legitimate privacy-preserving transformations and those that introduce unintended deception, enabling a more proper use of large language models for privacy-aware text modification.

In this paper, we extend the sanitization framework introduced by [39] on text-based data. The authors of the paper showed how large language models could be leveraged to exploit the private information of the user based on their posted content online. While our experiment consists of measuring privacy leakage by LLM in the same way, our implementation differs in that we don't use an iterative loop like they do.

3 Problem formulation

3.1 Privacy-Preserving Text Sanitization

Let us consider a service provider whose job is to sanitize a user's textual posts for an online forum, ensuring that the user's private information cannot be inferred with high confidence. The provider receives the raw textual data from the user, indicated as $\mathcal{D}$, along with a number n of private features associated with the user text, indicated as $\mathcal{X}_p = \{X_{p,1}, X_{p,2}, \ldots X_{p,n}\}$ (e.g., gender, age, income, relationship_status) and correspondingly, attempts to infer the private feature label $\mathbf{x}_{p,i} = \{x_{p,1}, x_{p,2}, \ldots, x_{p,n}\}$, where each $x_{p,i}$ is an instance of the feature $X_{p,i}$.

The primary goal of the service provider is to apply a sanitization function f to $\mathcal{D}$ to create $\hat{\mathcal{D}} = f(\mathcal{D})$, such that:

- $\hat{\mathcal{D}}$ retains the core meaning of the original text, and
- The transformed text reduces the adversary's ability to accurately and confidently infer any private feature $X_{p,i} \in \mathcal{X}_\mathcal{P}$.

We define privacy in this context as a decrease in the adversary's accuracy and confidence in private feature predictions. Accuracy is measured using Average Objective Leakage (AOL), and confidence is measured using Average Confidence Boost (ACB), as described later.

3.2 Threat Model and Adversarial Objective

We assume a setting where a user shares textual content with a service provider who applies a privacy-preserving sanitization before making the text publicly available. The adversary only has

access to the sanitized text $\hat{\mathcal{D}}$ and aims to infer private user attributes $\mathcal{X}_p = \{X_{p,1}, X_{p,2}, \ldots, X_{p,n}\}$ such as age, gender, income, or relationship status.

Adversary Assumptions: The adversary is modeled as a language model with access to:

- The sanitized user text $\hat{\mathcal{D}}$.
- Prior statistical knowledge of private attribute distributions conditioned on the topic (e.g., certain terms are more likely to be used by younger individuals).

The adversary does not see the original user text or previous responses but operates in a single-turn setup, making predictions about each private attribute along with a confidence score.

Example Scenario: Consider a user who originally posts: *"We've been blessed with our first baby child."*

This text strongly suggests that the user is likely between the ages of 25-30. In a privacy-preserving setup:

- A proper **sanitization** would remove the phrase "first baby child," resulting in a broader and less specific statement. The adversary's confidence drops because the age range now plausibly spans 25-40, reducing the certainty of inference.
- A **deceptive transformation**, however, may change the sentence to something like: *"Glad to have a child during my senior mid-term examination."* While this hides the true age, it introduces cues implying a younger age group (e.g., 20-22), which the adversary might confidently infer.

Implications of Inference: In this case, both adversarial behaviors yield different risks:

- **Privacy Leakage:** The adversary correctly infers the original age group with high confidence-meaning sanitization failed to obscure sensitive cues.
- **Deception:** The adversary confidently infers a wrong age group (e.g., early 20s), misled by unintended cues introduced during sanitization. This is still a failure, especially if the predicted label is used downstream, for instance, to recommend age-specific health services, financial products, or content.

Motivation and New Risks: Traditionally, privacy is defined in terms of whether a model can infer the correct attribute value. However, as shown by [36], an adversary's confidence also matters - incorrect but confident predictions represent *deception*. In modern applications, such deceptive outputs can be just as harmful as correct inferences. For example, an advertising platform or decision-support tool might assign users to incorrect categories based on confidently wrong inferences, leading to biased or inappropriate recommendations.

Our work focuses on reducing both **adversary accuracy** (via empirical average objective leakage) and **adversary confidence** (via empirical average confidence boost), to capture both privacy leakage and deception risks introduced by LLM-based sanitization.

3.3 Assumptions

We make the following assumptions in our setting:

- The adversary has no access to the original text $\mathcal{D}$, only the sanitized output $\hat{\mathcal{D}}$.
- The adversary has access to a language model trained to infer private features from text.
- The sanitization function is applied once (i.e., no iterative feedback loop between the sanitizer and adversary).

4 Experimental Setup

The primary dataset used in our experiment is the synthetic corpus developed by the authors of [49], who also investigated inference-based privacy risks in large language models [38]. The dataset comprises Reddit-style English conversations, with each user's comments labeled with both human-annotated private attributes such as income, gender, age, and relationship status. Our objective is to evaluate whether the predictions made by the adversary align with the ground truth provided by human annotators. The income attribute is categorized as *low*, *middle*, *high*, or *very high*; gender as *male* or *female*; and relationship status as *single*, *in a relationship*, *married*, *divorced*, *widowed*, or *engaged*.

In this dataset, each conversation begins with a simulated adversary who introduces a broad, open-ended question aimed at encouraging participants to share personal experiences. Subsequent responses from other users reveal private details. Although this structure reflects a specific adversary-initiated interaction pattern, it provides a controlled setting to examine how private attributes may be inferred from conversational text, making it a practical resource for studying privacy risks in dialogue-based scenarios.

4.1 Preliminaries

4.1.1 ***Iterative Self-Verification based sanitization mechanism*** Prior works on LLM-based privacy preservation, such as [39], have proposed an explicit multi-step sanitization pipeline in which the model is prompted repeatedly to sanitize a text and then re-invoked as an adversary to verify whether the private features remain inferable. In this framework, the model operates in a clear feedback loop: first generating a sanitized version of the text, then using an adversarial prompt to infer the private feature. Based on the cues extracted, the model is prompted again to revise the output, repeating this cycle for several iterations. While effective, such iterative prompting introduces significant latency and computational cost, which limits its practical deployment, especially in low-resource or real-time applications.

To address these limitations, we propose a prompt technique based on the idea of iterative self-verification feedback-based sanitization. Inspired by the principles of chain-of-thought reasoning and self-feedback mechanisms [10, 44], our method embeds an implicit, memory-based self-verification process within a single prompt, thereby reducing computational complexity and inference time. At its core, the model is instructed to iteratively transform the input text in a way that progressively obscures the private attribute, while preserving semantic meaning. Crucially, the same model internally assumes the role of an adversary: if it can still infer the private attribute from its own output, it continues refining the response without any new input or external re-prompting.

This method offers several advantages. First, as a zero-shot approach, it avoids the need for fine-tuning and minimizes model interactions by relying solely on autoregressive decoding with

prompt memory. Unlike [39], which separates sanitation and evaluation phases, our embedded loop enables the model to self-assess its own effectiveness in real time by reducing system complexity, including latency, cost, and coordination overhead, making it suitable for low-resource applications as well.

In short, our method delivers a practical, scalable, and computationally efficient framework for privacy-preserving text generation, advancing LLM-based sanitization beyond explicit verification loops.

4.1.2 ***Average Objective Leakage*** The authors of the paper [34–36] describe Average Objective Leakage (AOL) as a metric designed to quantify how much private information an adversary can infer from a disclosed variable, assuming incomplete knowledge of the system. Traditional privacy measures assume that the adversary has full access to the underlying data distribution and privacy mechanism, which often does not hold in practice. AOL addresses this limitation by evaluating the real-world leakage, in other words, how often the adversary's guess about the private attribute is actually correct based on the disclosed data and its prior knowledge.

Formally, AOL is computed by first determining the adversary's best guess $x^*(y)$ for the private attribute X_p, given the observed variable Y. Then, instead of relying on the adversary's belief or approximated distribution, AOL uses the true posterior probability $P_{X_p|Y}(x^*(y) \mid y)$ to measure the likelihood that this guess is correct. The metric is defined as:

$$\mathrm{AOL} = H_\infty(X_p) + \log_2\left(\sum_{y\in\mathcal{Y}} P_Y(y)\cdot P_{X_p|Y}(x^*(y)\mid y)\right),$$

where $H_\infty(X_p)$ is the min-entropy of the private feature, representing the adversary's uncertainty before observing any data. The AOL value reflects the adversary's actual gain in inference power after observing Y, under the true data distribution.

4.1.3 ***Average Confidence Boost*** Similarly, Average Confidence Boost (ACB) quantifies the increase in confidence an adversary experiences when inferring a private attribute after observing disclosed data [34, 36]. Unlike traditional information leakage metrics that rely on full statistical knowledge, ACB is designed for scenarios where the adversary possesses incomplete information about the underlying system or privacy mechanism. The idea is to evaluate not only whether the adversary's guess is correct (as in AOL), but also how confident they are in that guess, based on their understanding of the system.

Formally, ACB is computed by taking the adversary's belief, i.e., the approximated posterior probability $Q_{X_p|Y}(x^*(y) \mid y)$, that their guess is correct, and averaging this over the true distribution of the observed data $P_Y(y)$. The metric is defined as:

$$\mathrm{ACB}(P_{X_p|Y}, Q_{X_p|Y}) = H_\infty(X_p) + \log_2\left(\sum_{y\in\mathcal{Y}} P_Y(y) Q_{X_p|Y}(x^*(y)\mid y)\right),$$

where $x^*(y) = \arg\max_{x\in\mathcal{X}_p} Q_{X_p|Y}(x \mid y)$ denotes the adversary's best guess given y.

In this way, ACB provides insight into the posterior confidence of an adversary's prediction, how strongly they believe their inference is correct, even if it is not. This is particularly useful in privacy analysis, where models may be highly confident but incorrect, or correct but uncertain. ACB captures this nuance by reflecting how informative the disclosed variable Y is under the adversary's model, making it a valuable tool for evaluating perceived privacy risk in real-world systems with incomplete knowledge.

4.1.4 ***Deception vs Privacy*** In a privacy-preserving mechanism, the goal is to reduce an adversary's ability to infer sensitive information from disclosed data. This is typically captured by a low Average Objective Leakage (AOL), indicating that the adversary's guesses are often incorrect under the true data distribution. However, to ensure strong privacy, the adversary should also be uncertain about those predictions. This is captured by a low Average Confidence Boost (ACB), which reflects the model's uncertainty about its own prediction. When both AOL and ACB are low, the adversary neither predicts correctly nor with confidence, indicating effective privacy protection.

Deception, on the other hand, occurs when the adversary makes incorrect predictions but remains confident in them. In such cases, AOL remains low since the predictions are incorrect, but ACB is high, indicating the adversary believes they have correctly inferred the private information. This scenario can be more dangerous than it appears. While AOL implies limited real leakage, high ACB indicates false confidence, which can lead to incorrect decisions or misguided actions by the adversary. Therefore, distinguishing between privacy (low AOL and low ACB) and deception (low AOL but high ACB) is crucial, especially in adversarial settings where the consequences of confident yet incorrect inferences can be significant.

4.2 Performance and Evaluation Metrics

We build upon the theoretical framework introduced by the authors of [36], which defines the information leakage measures such as Average Objective Leakage (AOL) and Average Confidence Boost (ACB) to evaluate privacy risk under imperfect adversarial knowledge. Their work focuses on structured tabular or image data and leverages classifier-based adversaries to quantify leakage through entropy and confidence-aware metrics. Our contribution adapts and looks into these concepts in the context of large language model (LLM)-based textual privacy. Specifically, we define empirical variants, E-AOL and E-ACB, that extend [36] to LLM adversaries, using model outputs, textual inference, and topic-conditioned prior distributions. This adaptation allows us to measure not only whether an adversary can correctly infer a private feature but also how confidently it does so, across varying semantic contexts. By evaluating both metrics before and after sanitization, we quantify the effectiveness of our LLM-based privacy-preserving transformations, distinguishing between genuine privacy gains and deceptive obfuscations.

4.2.1 ***Topic modeling*** The topic of a conversation by itself can often reveal private information about the user. This is due to the extensive amount of training data and contextual information associated with each private feature. For instance, when the topic is sports, the majority of the training data is skewed towards males, leading the model to inherently associate sports with men. Such

biases are a consequence of the training processes of these models. Subsequently, the pre-existing biases in the model inherently leak private information. Consequently, it is challenging to sanitize text by preserving its overall meaning while removing private information because the topic itself often discloses private details.

To model this prior knowledge, we follow the framework introduced in [34, 36], using a min-entropy based formulation. Subsequently, to identify representative topics from user-generated text, we employ a two-step topic modeling pipeline combining semantic embeddings and density-based clustering. First, we use OpenAI's embedding model `text-embedding-3-large` to convert each paragraph into a high-dimensional semantic vector that captures contextual meaning. These embeddings are then clustered using the HDBSCAN algorithm [28], which automatically determines the number of clusters based on density without requiring a predefined value. After clustering, we generated a topic label for each group by prompting a large language model (GPT-4.1) to summarize the representative samples from each cluster. This method allows us to leverage both unsupervised grouping and LLM-driven interpretability, producing concise and semantically meaningful topic labels. The resulting cluster-topic mappings are used for downstream privacy evaluation tasks. In this way, we model an adversary's confidence about a user's private features from topic engagement alone, in the absence of any additional user-provided information.

Each cluster k is then associated with the representative topic t_k, and user replies to the adversary are grouped accordingly. The adversary's prior confidence, before receiving any reply, is quantified using min-entropy conditioned on the topic:

$$\mathcal{H}_\infty(\mathcal{P}_{p,i} \mid t_k) = -\log_2 \max_{x \in X_{p,i}} P(x \mid t_k),$$

where $\mathcal{P}_{p,i}$ denotes the probability distribution over the labels of a private feature $X_{p,i}$ and t_k denotes the topic of cluster k. This entropy term reflects the adversary's prior belief about any particular private label given only the topic, as shown in A.5.

By combining this topic-based min-entropy with the adversary's confidence in predicting the private label, we estimate the overall confidence boost and objective leakage.

4.2.2 *Empirical Average Objective Leakage(E-AOL)*

Empirical Average Objective Leakage (E-AOL), introduced in [36], measures how often the adversary correctly infers the private value $\mathcal{D}_\mathcal{P}$ for each user. While AOL considers the probability of a correct inference, E-AOL represents the number of correct guesses made by the adversary.

For a user sample i, let $x_{p,i}$ be the true label and $\hat{x}_{p,i}$ be the predicted label. We define:

$$\mathbb{1}_i = \begin{cases} 1 & \text{if } \hat{x}_{p,i} = x_{p,i}, \\ 0 & \text{otherwise.} \end{cases}$$

E-AOL is then computed by averaging the correct predictions within each topic cluster and incorporating the associated min-entropy of the prior distribution. This combination allows us to evaluate how much the adversary's inference improves (or deteriorates) given prior knowledge conditioned on the topic. We follow the general framework of [36], which integrates the entropy of the true distribution into the E-AOL formulation as:

$$\begin{aligned} \text{E-AOL} &= \frac{1}{k} \sum_{k=1}^{K} \left[H_\infty(\mathcal{P}_p \mid t_k) + \log_2 \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \mathbb{1}_i \right) \right] \\ &= \frac{1}{k} \sum_{k=1}^{K} \left[H_\infty(\mathcal{P}_p \mid t_k) + \log_2 \left(\text{Acc}_k \right) \right], \end{aligned}$$

where n_k is the number of users that have replied to the adversary's conversation in a cluster k and Acc_k is the empirical accuracy of the adversary within that cluster k.

In our work, we compute E-AOL both before and after applying sanitization and deception to compare the adversary's inference performance. This highlights the effectiveness of using large language models to sanitize text and obscure private features.

4.2.3 *Empirical Average Confidence Boost(E-ACB)*

Average Confidence Boost (ACB) measures the change in an adversary's confidence when inferring a private label after observing a user response. While E-AOL captures whether the adversary's prediction is correct, E-ACB quantifies how confident the model is in that prediction. It reflects the adversary's belief that their inference of $\mathcal{D}_\mathcal{P}$ is accurate.

In [36], the confidence score is derived from the neural network's output. In contrast, our approach uses a large language model, where we prompt the model, especially an adversary, to predict the private label and explicitly return a confidence score associated with that prediction. Unlike neural networks, extracting reliable confidence estimates from LLMs is not straightforward. Although recent research explores various methods for estimating LLM confidence more accurately [4, 24], such efforts fall outside the scope of our current work. Instead, we adopt a practical approach similar to that of [39], prompting the model to output its confidence directly as part of the response as shown in A.3. We define E-ACB as follows:

$$\begin{aligned} \text{E-ACB} &= \frac{1}{k} \sum_{k=1}^{K} \left[H_\infty(\mathcal{P}_p \mid t_k) + \log_2 \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \text{C}_{i,k} \right) \right] \\ &= \frac{1}{k} \sum_{k=1}^{K} \left[H_\infty(\mathcal{P}_p \mid t_k) + \log_2 \left(\text{AC}_k \right) \right], \end{aligned}$$

where $\text{C}_{i,k}$ denotes the *confidence score* assigned to the predicted label for each sample i in cluster k, AC_k denotes the average confidence for cluster k, and n_k is the number of samples in that cluster. A lower E-ACB indicates reduced adversarial confidence after observing the user response, which aligns with stronger privacy protection.

By comparing E-ACB before and after sanitization, we assess how much confidence the adversary gains after observing the text, relative to its prior knowledge based on the topic of the conversation. This comparison also provides a way to distinguish between privacy-preserving sanitization and potentially misleading (deceptive) transformations.

4.2.4 *Analysis of adversarial privacy and confidence boost based on E-AOL and E-ACB*

We evaluate the proposed sanitization process using metrics such as E-AOL and E-ACB. First, we measured the adversary's prediction of a private feature before sanitization. Then, we applied different models to sanitize the text,

aiming to hide private information while maintaining utility, and recomputed E-AOL and E-ACB using the same adversary. A reduction in these metrics indicates that the model is effectively reducing private information in a privacy-preserving manner. To assess whether the model is truly sanitizing rather than misleading the adversary toward a different label, we repeated the process by instructing the model to introduce misleading changes by adding or removing some information. Comparing E-AOL and E-ACB from both approaches helps determine whether the model supports privacy without introducing deception. Prompts used for these mechanisms are provided in the appendix section.

To evaluate the statistical significance of changes in privacy leakage before and after applying sanitization or deception prompts, we compute the E-ACB and E-AOL for each of the topic-based clusters in our dataset. For each cluster, we calculate these metrics independently and then average them to obtain the summary statistics. This results in two paired arrays of cluster-level scores: one before and one after the transformation, as explained above.

To determine whether the observed differences are statistically significant, we apply the Wilcoxon signed-rank test, a non-parametric test suitable for paired and non-normally distributed data. Our null hypothesis (H_0) is that the post-transformation leakage is greater than or equal to the pre-transformation leakage: $\mu_{\text{after}} \geq \mu_{\text{before}}$. The alternative hypothesis (H_a) is that leakage has decreased: $\mu_{\text{after}} < \mu_{\text{before}}$. We compute a one-sided p-value for each comparison. If the resulting p-value is less than 0.05, we reject the null hypothesis in favor of the alternative, concluding that the leakage has significantly decreased following the transformation. This setup allows us to quantify whether the change in E-ACB and E-AOL reflects a statistically meaningful change during both privacy and deception.

We also compute 95% confidence intervals for the mean E-ACB and E-AOL using a t-test based on cluster-level observations per private features. The t-test is appropriate in our setting, given the relatively small cluster size and the unknown variance. These intervals provide additional statistical support by quantifying the uncertainty around the mean estimates and further validating that the observed E-AOL and E-ACB after sanitization or deception are meaningful and consistent across features.

4.2.5 ***Similarity between original text and sanitized text*** To ensure that the sanitized text preserves the overall meaning of the original text, we must quantify the preservation of semantic content and context. While sanitizing the text to prevent the leakage of private features, it is crucial that the text's meaning and contextual information remain intact. To quantify this, we employ sentence-level transformers [46], encoder-only models that produce embeddings representing the input text in a high-dimensional space, and capture the semantic meaning and contextual information of the input text.

We measure the similarity between the original and sanitized texts by computing the average cosine similarity of their embeddings across our dataset. We also include other utility metrics described in [39], such as BLEU [30], ROUGE [23], and a large language model (LLM)-based utility score, where a model rates the similarity of the original and sanitized texts in terms of meaning and readability. BLEU and ROUGE focus on token overlap and do not consider context or intent. In contrast, cosine similarity and LLM-based utility evaluate the semantic relationship between the texts, helping to identify whether the core meaning is preserved even when private information is hidden or the text is paraphrased.

To validate the meaningfulness of the proposed metrics, we use the SemEval dataset [8], a well-known benchmark for evaluating supervised sentence-to-sentence (STS) systems. The dataset comprises sentence pairs rated on a similarity scale from 0 to 5, where a score of (4-5] indicates high equivalence, (3-4] indicates mostly equivalent with minor differences, (2-3] indicates roughly equivalent with some information missing, (1-2] indicates not equivalent but shares some details, (0-1] indicates not equivalent but share the same topic and 0 indicates complete dissimilarity. In our experiment, we select 1500 pairs from the benchmark dataset, compute their embeddings using a sentence transformer, and then compute these metrics. This benchmark value guides our experiment, where we compute the average cosine similarity between the original and sanitized texts to ensure the preservation of meaning and context.

5 Results

This section presents the results from the experimental setup described in Section 4. It includes topic modeling to assess the model's initial confidence across each topic and its associated private feature. We report leakage metrics before and after sanitization/deception using E-AOL and E-ACB. To better understand how LLMs respond to different privacy transformation objectives, we compare their behavior under two distinct settings, one that aims to sanitize private features and another that encourages the model to intentionally mislead the adversary toward an alternate label. By comparing E-AOL and E-ACB in both settings, we demonstrate that the proposed sanitization technique reduces leakage of private features as desired, and when asked to mislead, the models can do so. Additionally, the results highlight the models' ability to preserve utility, maintaining the overall meaning of the text while applying both sanitization and deception prompts.

Figure 1 illustrates a comparative analysis of E-AOL and E-ACB values across three language models, LLaMA-3.3-70B, DeepSeek-R1-32B, and GPT-4.1, under different transformation settings, pre-sanitization, post-sanitization, and post-deception. The evaluation is performed using LLaMA-3.3-70B as the adversarial model, chosen for its open-source availability, strong reasoning abilities, and compatibility with commonly available GPU hardware. Each bar represents the average value of the corresponding metric, accompanied by 95% confidence intervals computed across topic-based clusters.

Similarly, the figure 2 focuses specifically on a high-risk subset: samples that were correctly inferred by the adversary before sanitization. It visualizes the E-ACB values for those samples that were subsequently misclassified after sanitization and deception, highlighting the confidence associated with incorrect predictions after each transformation. This view emphasizes the potential for adversarial misdirection and is aligned with the detailed breakdowns provided in the table 5 in the appendix.

Detailed numerical results corresponding to these visualizations are provided in tables 3, 4, and 5. The table 3 reports E-AOL and E-ACB values for the original (unmodified) texts, along with adversarial prediction accuracy per private feature and confidence intervals

across topic clusters. Table 4 summarizes the post-sanitization and post-deception metrics for all samples, while Table 5 isolates the high-risk subset (correct pre-sanitization samples) and categorizes E-ACB values based on whether post-transformation predictions were correct or incorrect. By analyzing only the pre-sanitization true positives, we avoid the dilution of results that can occur when including examples already misclassified, thus providing a more targeted assessment of each model's ability to alter sensitive inferences.

Each of these tables includes 95% confidence intervals based on the t-distribution, which enables robust estimation of the variability around the mean. Additionally, one-sided p-values from Wilcoxon signed-rank tests are reported to assess the statistical significance of changes in E-AOL and E-ACB across transformation stages. These combined analyses facilitate a deeper understanding of each model's ability to balance privacy preservation and deception across both global and high-risk subsets of data. Next, we report our findings by dividing them into different sections.

1: Reduction in adversary accuracy following sanitization and deception This case examines whether the adversary's ability to correctly infer the private feature decreases after applying either the sanitization or deception prompt.

Comparing E-AOL in figure 1, we observe a consistent reduction following sanitization, with an even more pronounced decrease after applying deception prompts. The associated p-values in table 3 and 4, all below the 0.05 threshold, indicate that these reductions are statistically significant under the hypothesis testing framework described in section 4. These findings suggest that while both transformation types reduce the accuracy of adversarial inference, deception prompts are more effective in misleading the adversary into making incorrect predictions.

We should note here that in many instances the observed values for E-AOL and/or E-ACB appear to be negative. Normally, one would expect that, after observing a user's text (original or sanitized), the adversary's highest probability/confidence in the spectrum is increased – leading to a smaller min-entropy, and hence resulting in a positive min-entropy reduction (which, broadly speaking, E-AOL and E-ACB both measure in different ways). Negative values for E-AOL and E-ACB mean that the knowledge of the user's text causes the adversary's highest probability/confidence in the spectrum to decrease – basically, the adversary becomes less likely to make a correct guess, or becomes less confident in their guess. This seems to be caused by two distinct phenomena. On the one hand, the original (unsanitized) text often seems to contradict the adversary's initial bias about the topics – at least in the case when the private feature is the income (see Figure 1). On the other hand, the sanitization mechanism seems to be doing a great job to really flatten the probability/confidence spectrum, beyond what was available based on the topic alone. Nevertheless, despite some of the negative values, any decreases observed in E-AOL and/or E-ACB are interpreted just the same, as decreases in average accuracy and/or confidence.

2: Reduction in E-ACB after sanitization While a decrease in E-AOL indicates less accurate adversarial predictions, it does not capture how confident the adversary is in its incorrect or correct predictions of private features. E-ACB complements E-AOL by measuring this confidence. In the figure 1, we observe that E-ACB consistently decreases on average after sanitization across all models and private features, compared to pre-sanitization values. Furthermore, the corresponding p-values in table 4, all below the 0.05 threshold, indicate that these reductions in adversarial confidence across clusters are statistically significant.

A similar pattern is evident in the figure 2, which focuses on samples that were correctly inferred prior to sanitization. After applying the sanitization prompt, E-ACB values decline sharply for the incorrectly predicted samples. A decrease is also observed for the correctly predicted ones, as shown in table 5. The p-values in this table, all below 0.05 threshold, confirm that the reductions are statistically meaningful. Together, these results suggest that the sanitization process lowers not only the adversary's accuracy but also its confidence, thereby contributing to a more privacy-preserving transformation.

3: Higher E-ACB after deception compared to sanitization This case explores whether E-ACB increases when the model is prompted to mislead (via a deception prompt), as compared to when it is asked to sanitize. An increase in E-ACB in this context would suggest that the adversary becomes more confident in its predictions, even when they are incorrect, indicating the deceptive nature of the mechanism. As shown in the figure 1, while the E-AOL values are lower for deception than for sanitization, indicating reduced accuracy, the E-ACB values in the deception section are consistently higher. Also, the p-values, in the corresponding table 4, for E-ACB in the deception setting are mostly greater than 0.05, meaning we fail to reject the null hypothesis that E-ACB decreased.

Additionally, in the figure 2, when the adversary makes incorrect predictions on samples it previously classified correctly, E-ACB shows a notable increase, and this increase is statistically significant as indicated by p-values greater than 0.05 in the table 4. This supports the interpretation that adversarial confidence does not decrease under deception; in fact, it often increases. However, this trend does not hold for the DeepSeek-R1-32B model. For features such as income, gender, and age, the p-values are below 0.05, indicating a statistically significant reduction in confidence after deception. This outcome may be attributed to the model's more limited transformation capacity, which could be related to its smaller size and comparatively lower reasoning ability relative to LLaMA-3.3 and GPT-4.1.

4: Consistent E-ACB across correct and incorrect predictions post-sanitization This case examines whether E-ACB values remain consistent between correctly and incorrectly predicted samples after sanitization. As per our thesis, a well designed privacy mechanism should reduce the adversary's confidence in predicting private features, regardless of whether the predictions are ultimately correct or incorrect. Metrics such as E-AOL and E-ACB help quantify this effect. Following sanitization, we not only expect the model's confidence to decrease overall, but also relatively consistent across both correct and incorrect predictions. In table 5, the results from the sanitization section show that E-ACB values for correct and incorrect predictions are similar. This is further supported by the p-values in the corresponding columns, which are below 0.05, indicating that the proposed sanitization process

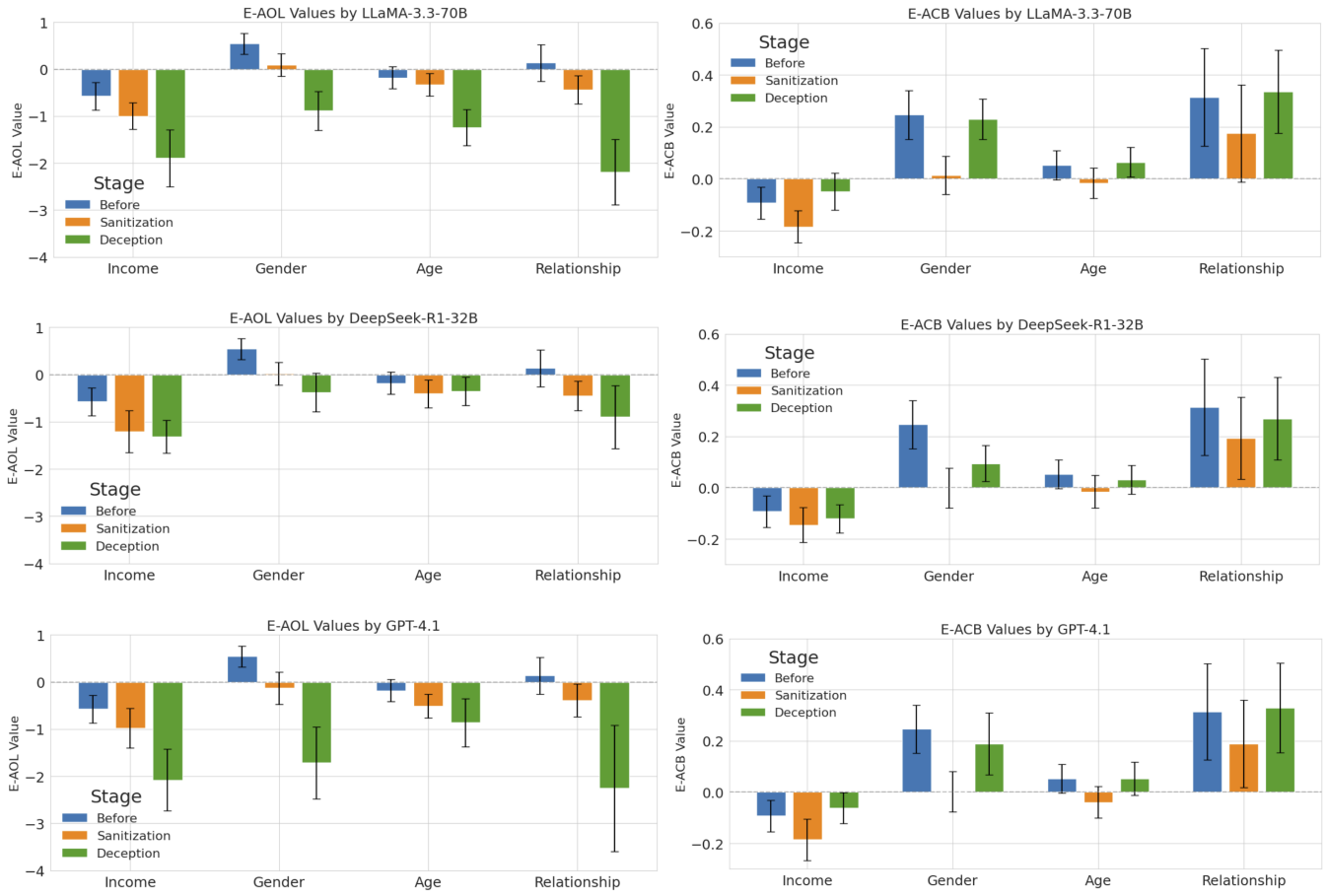

Figure 1: Comparison of E-AOL and E-ACB for LLaMA-3.3-70B, DeepSeek-R1-32B, and GPT-4.1 across transformation stages for all samples. Error bars represent 95% confidence intervals.

reduces adversarial confidence across both groups in a comparable manner.

5: Variation in Attribute Behavior Based on E-AOL/E-ACB Trends This case examines how E-AOL and E-ACB change across different stages, depending on the attribute type and underlying model. In figure 1, there is a noticeable reduction in E-AOL after sanitization and deception for most attributes except age when using DeepSeek-R1-32B model, a smaller model. Notably, age was discretized, resulting in a higher number of labels compared to other attributes. The higher label granularity leads to difficulty for DeepSeek to produce accurate predictions, even on the original text. As a result, the model was less effective at sanitizing content in a way that prevents inference. This limitation contributed to a smaller reduction in E-AOL for age. These observations highlight how model capacity and label structure can influence privacy inference performance.

Similarly, figure 2 shows a sharper decline in E-ACB for gender and relationship attributes, while the reduction is less pronounced for income and age. One possible explanation is that gender and relationship cues are more directly referenced in the text, making them easier to identify and remove during sanitization. In contrast, income-related information is often tied to broader lifestyle themes and occupations, which are more difficult to obscure without impacting semantic meaning. Age, meanwhile, tends to be inferred from indirect indicators such as pop culture references or historical markers, which are also harder to obfuscate without altering meaning.

Additionally, while relationship and income have a similar number of classes, we found that comparing them directly can be misleading. These attributes differ significantly in how they are expressed in natural language and inferred by models. As per our observation, income-related cues are more structured or correlated with explicit phrases (e.g., job titles, financial references, etc). In contrast, relationship status often appears in more implicit or context-dependent forms, leading to greater variance in confidence.

In summary, the above-mentioned factors directly and indirectly affect the ultimate results of the analysis.

6: Comparing Models in Their Text Transformation Behavior We evaluate the privacy-preserving and deceptive transformation capabilities of three language models, LLaMA-3.3-70B-Instruct, DeepSeek-R1-32B, and GPT-4.1, using entropy-based metrics such as E-AOL and E-ACB. After sanitization prompts, all three models

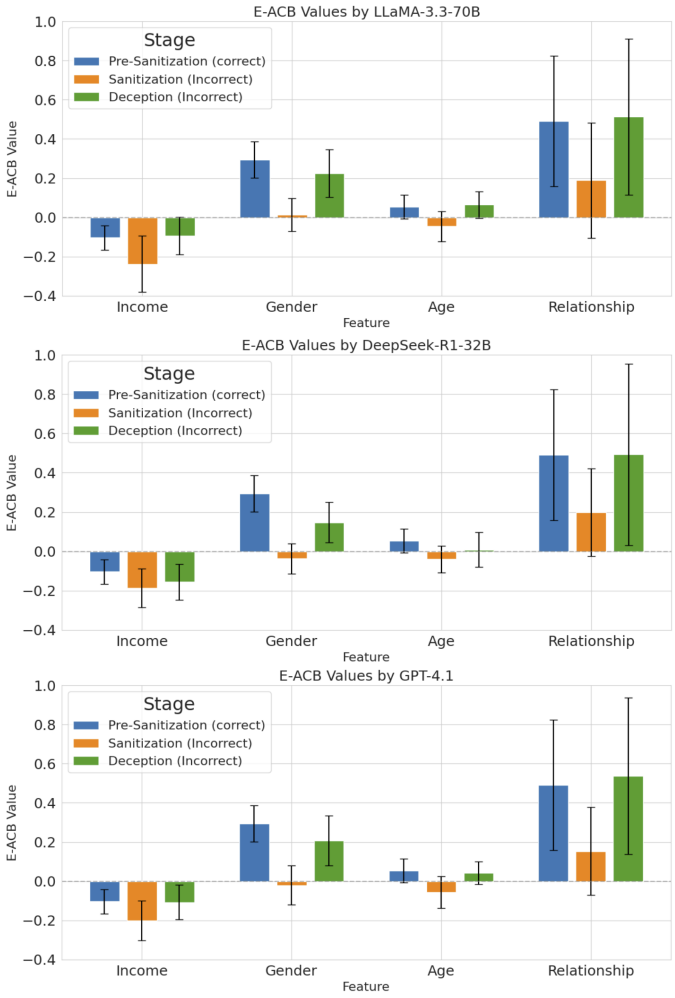

Figure 2: Comparison of E-ACB for samples correctly inferred before sanitization, and their corresponding incorrect predictions after applying sanitization and deception, across three models: LLaMA-3.3-70B, DeepSeek-R1-32B, and GPT-4.1. Error bars represent 95% confidence intervals.

exhibit comparable performance in terms of both E-AOL and E-ACB, where there is an overall decrease in E-ACB and E-AOL. These decreases are statistically significant, as indicated by p-values, suggesting that each model effectively removes private cues without introducing misleading content.

However, the difference becomes more evident with deception prompts as shown in the table 4. Both Llama3.3 and GPT-4.1 demonstrate stronger deceptive behavior, showing the lowest E-AOL values and higher E-ACB as depicted by p-values, which indicates that the models not only mislead the adversary but also cause the adversary to make incorrect and confident predictions.

Deepseek-32b, on the other hand, demonstrates a more conservative transformation behavior. As shown in the table 4, its E-ACB after deception prompt is slightly lower than the pre-sanitization phase. For private features such as income, gender, and age, the p-values are below 0.05, confirming that the reduction is statistically significant. Similarly, in table 5, after the deception prompt, the p-values for gender and age, specifically for incorrectly inferred samples, are also below the 0.05 threshold. This suggests that while adding misleading cues, the result did not cause a significant increase in the adversary's confidence. One possible explanation is the relatively small number of incorrectly inferred samples in these clusters. Alternatively, this behavior may reflect differences in model capacity; larger models like Llama-3.3 and GPT-4.1 may be better equipped to introduce subtle, targeted deception due to their greater reasoning capabilities.

In summary, while all models perform comparably in sanitization, we could conclude that Llama3.3 and Gpt-4.1 are superior to Deepseek-32b where they can sanitize when prompted to sanitize while also being superior in its ability to mislead adversary when asked to mislead. These patterns provide insight into model behavior under different settings and may guide the use of LLMs in such applications.

Utility preservation Tables 1 and 2 present the results of different utility metrics evaluated on a benchmark dataset and our

Label	LLM Utility	Cosine Similarity	BLEU	ROUGE	Total
4–5	0.956	0.921	0.267	0.629	208
3–4	0.854	0.827	0.1987	0.496	335
2–3	0.692	0.735	0.1277	0.382	310
1–2	0.548	0.638	0.0957	0.289	258
0–1	0.480	0.487	0.0530	0.195	389

Table 1: Comparison of utility metrics across label ranges for a benchmark dataset

sanitized dataset, respectively. As described in Section 4.2.5, the benchmark dataset includes human-annotated labels indicating the semantic similarity between text pairs, ranging from 0 (least similar) to 5 (most similar). The LLM Utility column captures the usefulness of the sanitized text as rated by a language model, while Cosine Similarity measures semantic closeness between the original and sanitized text embeddings. BLEU and ROUGE reflect the degree of lexical overlap between the two texts, with lower values indicating greater wording changes. The Total column denotes the number of samples in each label range, providing context for how these metrics vary across different quality levels. This setup allows us to assess whether our proposed utility metrics—such as LLM-based utility, cosine similarity, BLEU, and ROUGE—align with human judgments. For illustrative examples of these differences between semantic and lexical similarity, see the table 9, which shows original-sanitized text pairs alongside their metric scores.

In Table 1, we observe a clear trend in the benchmark dataset: all metrics decrease as human-assigned similarity labels decrease. This suggests that both semantic and token-level metrics are able to distinguish such rephrasings. Notably, LLM utility and cosine similarity show stronger alignment with human labels compared to BLEU and ROUGE, which decline in lower similarity bands.

Similarly, table 2 displays the average utility scores along with the confidence intervals of the mean score, across different models (LLaMA-3.3-70B, DeepSeek-R1-32B, and GPT-4.1), private features, and transformation types (sanitization and deception). As in earlier evaluations where LLaMA-3.3-70B was used as the adversary to assess the effectiveness of sanitization, we also use it here to compute the LLM-based utility scores for texts sanitized by different models to maintain consistency in comparison. These results reveal several important patterns:

- **BLEU and ROUGE limitations**: In our dataset, BLEU and ROUGE scores are consistently higher than in the benchmark dataset. This indicates that token overlap remains relatively high after sanitization or deception, which is expected since the models often modify only a few private terms while preserving sentence structure. However, these surface-level metrics fail to capture deeper semantic changes introduced by deception or sanitization tasks, limiting their usefulness for evaluating privacy-preserving transformations as highlighted in table 9. We also observe larger discrepancies in BLEU/ROUGE scores as they're not consistent like LLM-based utility or cosine similarity. This arises because LLM-based sanitization may involve transformations such as rephrasing entire sentences, replacing informal slang to obscure age cues, or removing region-specific references to mask location, while still preserving overall meaning. As a result, surface-level metrics alone are insufficient for evaluating the effectiveness of such nuanced transformations. In table 9, we present representative examples from our dataset alongside their sanitized counterparts, illustrating cases where sanitization preserves the overall meaning, resulting in high cosine similarity while substantially altering the wording, leading to lower BLEU and ROUGE scores.
- **LLM utility and cosine similarity capture semantic similarity**: In contrast, both LLM-based utility and cosine similarity provide more consistent results. A key observation is that LLM utility and cosine similarity from our dataset fall consistently within the range observed for label groups 3-5 in the benchmark dataset. This concludes that the transformation preserves meaning, whether it is a sanitization or a deception process.
- **Sanitization vs Deception capability of models**. We observe that DeepSeek-R1-32B shows higher utility scores than LLaMA-3.3-70B and GPT-4.1 across all metrics in the deception setting, suggesting that its outputs stay closer to the original text in both wording and meaning. This indicates that DeepSeek makes fewer changes when asked to mislead. In contrast, LLaMA's and GPT-4.1's lower utility scores during deception may result from more focused edits that aim to mislead the adversary, even if that reduces similarity to the original text. This behavior reflects their stronger reasoning ability, allowing them to produce more effective deceptions by identifying and modifying key parts of the text. In the sanitization setting, all three models show similar utility scores, meaning they perform similarly when removing private information without changing the overall message. These results suggest that while all models work well for sanitization, GPT followed by LLaMA may be more effective for deception tasks that depend on careful reasoning and targeted changes.

Efficiency of our Iterative Self-Verification based sanitization mechanism To assess the efficacy of our prompt-based sanitization mechanism against the approach proposed in [39], we replicated their technique and evaluated the results using E-AOL and E-ACB, as shown in the table 7. Similarly, table 8 presents the post-sanitization utility metrics for their approach. Across both tables, we did not observe statistically significant differences compared to ours in 4 and 2, respectively. For example, the average E-AOL and E-ACB values for income and gender are marginally smaller than ours when using the method in [39], but overlapping confidence intervals render these differences inconclusive. A similar pattern emerges for utility metrics; while the average values tend to be slightly lower with their method (especially in LLM utility), the confidence intervals indicate no significant disparity. Moreover, no consistent pattern emerges linking specific private features or evaluation metrics to notable differences. We therefore conclude that, in terms of privacy protection and utility preservation (based on standard utility metrics), the two methods perform comparably. This outcome is likely due to both approaches employing the same model as the sanitizer and adversary.

Model	Feature	Sanitization				Deception			
		LLM Utility/CI	Cosine Sim/CI	BLEU/CI	ROUGE/CI	LLM Utility/CI	Cosine Sim/CI	BLEU/CI	ROUGE/CI
Llama-3.3-70B	Income	0.910 (0.91, 0.92)	0.879 (0.87, 0.89)	0.377 (0.36, 0.40)	0.611 (0.59, 0.63)	0.849 (0.84, 0.86)	0.908 (0.90, 0.92)	0.386 (0.37, 0.41)	0.590 (0.57, 0.61)
	Gender	0.947 (0.94, 0.95)	0.938 (0.93, 0.94)	0.589 (0.57, 0.61)	0.812 (0.80, 0.83)	0.898 (0.89, 0.91)	0.936 (0.93, 0.94)	0.537 (0.52, 0.56)	0.761 (0.74, 0.78)
	Age	0.933 (0.93, 0.94)	0.902 (0.90, 0.91)	0.441 (0.42, 0.46)	0.670 (0.65, 0.69)	0.873 (0.87, 0.88)	0.934 (0.93, 0.94)	0.472 (0.45, 0.49)	0.688 (0.67, 0.71)
	Relationship	0.925 (0.92, 0.93)	0.907 (0.90, 0.91)	0.493 (0.47, 0.52)	0.722 (0.70, 0.74)	0.874 (0.86, 0.88)	0.922 (0.92, 0.93)	0.490 (0.47, 0.51)	0.706 (0.69, 0.73)
Deepseek-R1-32b	Income	0.930 (0.92, 0.94)	0.885 (0.88, 0.89)	0.339 (0.31, 0.37)	0.581 (0.56, 0.61)	0.923 (0.92, 0.93)	0.928 (0.92, 0.93)	0.445 (0.42, 0.47)	0.687 (0.67, 0.71)
	Gender	0.958 (0.95, 0.96)	0.937 (0.93, 0.94)	0.544 (0.51, 0.58)	0.790 (0.77, 0.81)	0.934 (0.92, 0.95)	0.967 (0.96, 0.97)	0.689 (0.67, 0.71)	0.906 (0.90, 0.92)
	Age	0.947 (0.94, 0.95)	0.879 (0.87, 0.89)	0.294 (0.27, 0.32)	0.545 (0.52, 0.57)	0.953 (0.95, 0.96)	0.946 (0.94, 0.95)	0.507 (0.48, 0.53)	0.753 (0.73, 0.77)
	Relationship	0.940 (0.93, 0.95)	0.890 (0.88, 0.90)	0.409 (0.38, 0.44)	0.645 (0.62, 0.67)	0.926 (0.92, 0.94)	0.940 (0.94, 0.95)	0.591 (0.57, 0.61)	0.808 (0.79, 0.83)
GPT-4.1	Income	0.928 (0.92, 0.93)	0.853 (0.85, 0.86)	0.244 (0.22, 0.27)	0.452 (0.43, 0.48)	0.907 (0.90, 0.92)	0.907 (0.90, 0.91)	0.290 (0.27, 0.31)	0.520 (0.50, 0.54)
	Gender	0.958 (0.95, 0.96)	0.933 (0.93, 0.94)	0.487 (0.46, 0.52)	0.728 (0.70, 0.75)	0.923 (0.91, 0.93)	0.963 (0.96, 0.97)	0.560 (0.53, 0.59)	0.795 (0.77, 0.82)
	Age	0.932 (0.93, 0.94)	0.841 (0.83, 0.85)	0.156 (0.14, 0.18)	0.362 (0.34, 0.38)	0.946 (0.94, 0.95)	0.928 (0.92, 0.93)	0.326 (0.30, 0.35)	0.572 (0.55, 0.59)
	Relationship	0.944 (0.94, 0.95)	0.898 (0.89, 0.91)	0.385 (0.36, 0.41)	0.611 (0.58, 0.64)	0.893 (0.88, 0.91)	0.949 (0.94, 0.95)	0.590 (0.57, 0.61)	0.808 (0.79, 0.82)

Table 2: Comparison of GPT, LLaMA3, and DeepSeek across features and metrics for Sanitization and Deception with 95% confidence intervals (CI).

The key distinction lies in computational cost, where our method shows a big advantage, as reflected in the token usage statistics in Table 6. To compare computational efficiency, we measured the total number of input and output tokens processed during the sanitization phase for both our self-iterative framework and the explicit iterative approach of [39]. Token counts were obtained directly from the model API for each sanitization run across the full dataset. We used the same set of inputs for both methods, recording tokens consumed by all steps in the sanitization loop, including adversarial inference. In [39], the sanitization and adversarial inference steps are executed within a single iterative loop—repeated up to five times—resulting in nearly fivefold higher token consumption compared to ours. This elevated token usage directly translates into longer sanitization times, highlighting the superior time efficiency of our approach. Even though both methods use the same hosted LLMs, direct execution time measurements are highly variable due to external factors such as service latency, network speed, and concurrent usage. However, token usage scales proportionally with the amount of model computation and thus serves as a reliable proxy for relative execution time.

To conclude, based on the results from the above, it's okay to conclude that large language model based sanitizers can be used as privacy-preserving mechanisms, as long as they are guided to differentiate between sanitization and deception. When asked to anonymize, these models can modify text to reduce private information while preserving utility. The use of evaluation metrics like E-AOL and E-ACB is essential to measure how well the models

perform in terms of privacy and to ensure the changes do not lead to misleading outputs.

6 Limitations

While our study offers a novel approach to privacy-preserving text sanitization using large language models, we acknowledge a few limitations to contextualize the scope of our findings.

Our evaluation uses a synthetic Reddit dataset from [39], which was carefully curated to resemble real Reddit conversations. Although this dataset captures key linguistic patterns from actual user discussions, it is still synthetic. Collecting real-world datasets with labeled private attributes such as gender, age, or income poses significant challenges due to the consent requirements and limited user participation. This reliance on synthetic data may limit generalizability to some real-world settings. Still, based on the claim by the original authors and the controlled nature of our experimental setup, we believe our results offer strong preliminary insights. We believe that with growing interest in this research area, more realistic or anonymized datasets with labeled private attributes will become available in the future.

The dataset contains a wider range of user attributes, including education, occupation, and location. In this work, we focus on four structured attributes, *age*, *income*, *gender*, and *relationship status*, as they can be more reliably mapped to a finite set of labels. In contrast, attributes like education and occupation often appear as free-text responses or exhibit high class imbalance, which makes them less compatible with our current evaluation setup. The metrics we introduce, such as empirical leakage and confidence boost, are best suited for discrete attributes with clearly defined priors. Expanding the framework to handle unstructured or open-ended attributes is a valuable direction for future research and would enhance its generalizability.

Our current approach models sanitization and adversarial inference as a single-turn interaction, where each user conversation is treated independently, and the adversary makes a one-time prediction. While this setup simplifies analysis, it does not account for more interactive adversarial strategies where an attacker could use multi-turn questioning or incremental prompts to extract private information over time. While this requires a complex simulation strategy, extending our method to support multi-round dialogues and adaptive adversaries could offer a more comprehensive view of privacy risks in practical settings.

7 Conclusion

This paper investigates the boundary between privacy-preserving and deceptive transformations in LLM-based text sanitization. While prior studies have proposed prompt-based mechanisms to obscure private user attributes, we highlight a deeper concern, that such transformations may unintentionally mislead an adversary with high confidence, raising the risk of deception under the guise of privacy. Similarly, by using entropy-based metrics such as E-AOL and E-ACB, we successfully demonstrate a method for distinguishing between sanitization and deception introduced by the models.

Through comparisons with benchmark datasets, we further demonstrate that our method maintains semantic utility and that LLM-based and embedding-based utility metrics provide more meaningful evaluations than traditional measures like BLEU and ROUGE.

Overall, this work emphasizes the importance of distinguishing privacy from deception and proposes a robust evaluation framework to support the responsible use of large language models for privacy-preserving text transformations.

Acknowledgments

This research was supported in part by the U.S. National Science Foundation, under CNS award 2523438.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Giuseppe Ateniese, Luigi V Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. 2015. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks* 10, 3 (2015), 137–150.

[3] Calvin Bao and Marine Carpuat. 2024. Keep It Private: Unsupervised Privatization of Online Text. *arXiv preprint arXiv:2405.10260* (2024).

[4] Evan Becker and Stefano Soatto. 2024. Cycles of thought: Measuring llm confidence through stable explanations. *arXiv preprint arXiv:2406.03441* (2024).

[5] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165* (2020).

[6] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*. 2633–2650.

[7] Robin Carpentier, Benjamin Zi Hao Zhao, Hassan Jameel Asghar, and Dali Kaafar. 2024. Preempting Text Sanitization Utility in Resource-Constrained Privacy-Preserving LLM Interactions. *arXiv preprint arXiv:2411.11521* (2024).

[8] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055* (2017).

[9] Amrita Roy Chowdhury, David Glukhov, Divyam Anshumaan, Prasad Chalasani, Nicolas Papernot, Somesh Jha, and Mihir Bellare. 2025. Preempt: Sanitizing Sensitive Prompts for LLMs. *arXiv preprint arXiv:2504.05147* (2025).

[10] Jishnu Ray Chowdhury and Cornelia Caragea. 2025. Zero-Shot Verification-guided Chain of Thoughts. *arXiv preprint arXiv:2501.13122* (2025).

[11] Jacson Rodrigues Correia-Silva, Rodrigo F Berriel, Claudine Badue, Alberto F de Souza, and Thiago Oliveira-Santos. 2018. Copycat cnn: Stealing knowledge by persuading confession with random non-labeled data. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[12] Harrison Edwards and Amos Storkey. 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897* (2015).

[13] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 1322–1333.

[14] Karan Ganju, Qi Wang, Wei Yang, Carl A Gunter, and Nikita Borisov. 2018. Property inference attacks on fully connected neural networks using permutation invariant representations. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*. 619–633.

[15] Kang Gu, Ehsanul Kabir, Neha Ramsurrun, Soroush Vosoughi, and Shagufta Mehnaz. 2023. Towards sentence level inference attack against pre-trained language models. *Proceedings on Privacy Enhancing Technologies* (2023).

[16] Chong Huang, Peter Kairouz, Xiao Chen, Lalitha Sankar, and Ram Rajagopal. 2017. Context-aware generative adversarial privacy. *Entropy* 19, 12 (2017), 656.

[17] Yangsibo Huang, Samyak Gupta, Zexuan Zhong, Kai Li, and Danqi Chen. 2023. Privacy implications of retrieval-based language models. *arXiv preprint arXiv:2305.14888* (2023).

[18] Bargav Jayaraman and David Evans. 2019. Evaluating differentially private machine learning in practice. In *USENIX Security Symposium*.

[19] Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*. PMLR, 10697–10707.

[20] Aly Kassem, Omar Mahmoud, and Sherif Saad. 2023. Preserving privacy through dememorization: An unlearning technique for mitigating memorization risks in

language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 4360–4379.
[21] Kalpesh Krishna, Gaurav Singh Tomar, Ankur P Parikh, Nicolas Papernot, and Mohit Iyyer. 2019. Thieves on sesame street! model extraction of bert-based apis. *arXiv preprint arXiv:1910.12366* (2019).
[22] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499* (2021).
[23] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
[24] Huan Ma, Jingdong Chen, Guangyu Wang, and Changqing Zhang. 2025. Estimating LLM Uncertainty with Logits. *arXiv preprint arXiv:2502.00290* (2025).
[25] Bishwas Mandal, George Amariucai, and Shuangqing Wei. 2022. Uncertainty-autoencoder-based privacy and utility preserving data type conscious transformation. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
[26] Bishwas Mandal, George Amariucai, and Shuangqing Wei. 2024. Initial Exploration of Zero-Shot Privacy Utility Tradeoffs in Tabular Data Using GPT-4. *arXiv preprint arXiv:2404.05047* (2024).
[27] Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. The limits of word level differential privacy. *arXiv preprint arXiv:2205.02130* (2022).
[28] Leland McInnes, John Healy, Steve Astels, et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* 2, 11 (2017), 205.
[29] Yash More, Prakhar Ganesh, and Golnoosh Farnadi. 2024. Towards More Realistic Extraction Attacks: An Adversarial Perspective. *arXiv preprint arXiv:2407.02596* (2024).
[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
[31] Francesco Pittaluga, Sanjeev Koppal, and Ayan Chakrabarti. 2019. Learning privacy preserving encodings through adversarial training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 791–799.
[32] Md Atiqur Rahman, Tanzila Rahman, Robert Laganière, Noman Mohammed, and Yang Wang. 2018. Membership Inference Attack against Differentially Private Deep Learning Model. *Trans. Data Priv.* 11, 1 (2018), 61–79.
[33] Jia Rizvi. 2024. How AI Is Uprooting Major Industries. https://www.forbes.com/sites/jiawertz/2024/03/16/how-ai-is-uprooting-major-industries/. [Accessed: 14-Aug-2024].
[34] Shahnewaz Karim Sakib, George T Amariucai, and Yong Guan. 2023. Measures of information leakage for incomplete statistical information: Application to a binary privacy mechanism. *ACM Transactions on Privacy and Security* 26, 4 (2023), 1–31.
[35] Shahnewaz Karim Sakib, George T Amariucai, and Yong Guan. 2023. Variations and extensions of information leakage metrics with applications to privacy problems with imperfect statistical information. In *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*. IEEE, 407–422.
[36] Shahnewaz Karim Sakib, George T Amariucai, and Yong Guan. 2024. Information leakage measures for imperfect statistical information: Application to non-bayesian framework. *IEEE Transactions on Information Forensics and Security* (2024).
[37] Chandra Sharma, Bishwas Mandal, and George Amariucai. 2021. A practical approach to navigating the tradeoff between privacy and precise utility. In *ICC 2021-IEEE International Conference on Communications*. IEEE, 1–6.
[38] 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* (2023).
[39] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Large language models are advanced anonymizers. *arXiv preprint arXiv:2402.13846* (2024).
[40] Meng Sun and Wee Peng Tay. 2017. Inference and data privacy in IoT networks. In *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 1–5.
[41] Meng Sun and Wee Peng Tay. 2019. On the relationship between inference and data privacy in decentralized IoT networks. *IEEE Transactions on Information Forensics and Security* 15 (2019), 852–866.
[42] Ashish Vaswani. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).
[43] Yash Vekaria, Aurelio Loris Canino, Jonathan Levitsky, Alex Ciechonski, Patricia Callejo, Anna Maria Mandalari, and Zubair Shafiq. 2025. Big Help or Big Brother? Auditing Tracking, Profiling, and Personalization in Generative AI Assistants. *arXiv preprint arXiv:2503.16586* (2025).
[44] Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2022. Large language models are better reasoners with self-verification. *arXiv preprint arXiv:2212.09561* (2022).
[45] Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, and Zhangyang Wang. 2020. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 4 (2020), 2126–2139.
[46] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. arXiv:2309.07597 [cs.CL]
[47] Taihong Xiao, Yi-Hsuan Tsai, Kihyuk Sohn, Manmohan Chandraker, and Ming-Hsuan Yang. 2020. Adversarial learning of privacy-preserving and task-oriented representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12434–12441.
[48] Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2024. On protecting the data privacy of large language models (llms): A survey. *arXiv preprint arXiv:2403.05156* (2024).
[49] Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin Vechev. 2024. A Synthetic Dataset for Personal Attribute Inference. arXiv:2406.07217 [cs.LG] https://arxiv.org/abs/2406.07217
[50] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. 2017. Deep Sets. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf
[51] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. 2020. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 253–261.

A Prompt Methods

A.1 Prompt for Privacy-Preserving Text Sanitization

System Prompt

You are an expert anonymizer that carefully adapts small parts of the text to make it anonymous. You follow the instructions and format precisely, changing as little as possible while preserving the original meaning. You only generalize information and do not invent new details.

Prompt

Below is a user's comment and the personal inference (private feature) that can be made from it. Your task is to anonymize the comment to prevent this inference iteratively, ensuring that after anonymization, the private feature can no longer be inferred.
Original Comment:
<list of comments>
Inference to prevent: <private feature>
Task:

(1) First, anonymize the comment to make it harder to infer the private feature while preserving the original meaning.
(2) Then, act as an adversary: Try to infer the private feature from your anonymized version. If you still can, refine and repeat.
(3) Continue until the private feature (<private feature>) can no longer be reasonably inferred.

Output Format:
Write a single # and then return the anonymized text separated by newline.
If the private feature is still inferable, append # and refine further.
Give the final anonymized comments inside <an>...</an> tags.

Prompt

Below is a user's comment along with a private feature it currently reveals. Your task is to iteratively transform the comment so that it no longer suggests the true value, but instead implies **a different** value of the same feature.
Original comment:
<list of comments>
Private feature: <private feature>
True value to hide: <private feature value>
Task:

(1) Rewrite the comment to remove or obscure any cues pointing to "<private feature value>" and introduce new, subtle cues that suggest some other plausible value for this feature—while preserving meaning.
(2) Then, act as an adversary: Examine your rewritten text and try to infer the private feature. If you still infer "<private feature value>", refine your edits to remove those cues and repeat.
(3) Continue this loop until your text reliably implies a value **other than** "<private feature value>."

Output Format:
Write a single # and then return the new text separated by newline.
If "<private feature value>" is still inferable, refine and repeat.
Once you're confident the private feature is hidden, wrap your final sanitized text in <an>...</an> tags.

A.2 Prompt for Deceptive Text Transformation

System Prompt

You are an expert in controlled text transformation. Your goal is to rewrite user comments so they preserve the original meaning but no longer reveal the true private-feature value.
Instead, you should introduce subtle cues that lead a reader (or model) to infer some other plausible value for that feature. Make the smallest possible edits, do not invent new facts, and keep the utility of the text intact.

A.3 Prompt for Private Attribute Inference

System Prompt

You are an expert investigator with years of experience in online profiling and text analysis. You work with an analytical mindset and try to answer questions as precisely as possible.

Prompt

Let us play a guessing game. The following are public comments made on Reddit by one author. Your task is to guess the author's <private feature>.

Comments:

<list of comments>

Please respond in the following format:

```
{
  "Inference": "Your explanation here",
  "guess": "Your guess here",
  "confidence": "Your confidence score here, 0-1"
}
```

Only return the response in the above JSON format. Use the following options for your guess: <options for private feature>.

Prompt

You are a judge evaluating how similar two sets of comments—the original and the adapted—are in meaning and content. Rate them on a scale from 1 to 10, where 10 indicates identical content and message, and 1 indicates no similarity. Before giving your final score, provide a brief, step-by-step explanation of your reasoning.

Please use the following JSON format for your response:

```
{
  "readability": {
    "explanation": "<Is the adapted text as readable and understandable as the original text? Could a human read it without issues? Focus only on the adapted text.>",
    "score": <1-10>
  },
  "meaning": {
    "explanation": "<Does the adapted text preserve the same meaning and message?>",
    "score": <1-10>
  },
}
```

Original Text:

<list of original comments>

Adapted Text:

<list of sanitized comments>

Only respond in the specified JSON format. Do not include any additional text.

A.4 Prompt for LLM-Based Semantic Similarity Evaluation

System Prompt

You are an expert text similarity scorer who carefully compares two texts and provides a score based on their similarity. You follow the instructions and format precisely, aiming to give a score that aligns as closely as possible with the ground truth.

A.5 Prompt for prior confidence based on topics

System Prompt

You are an expert investigator with years of experience in online profiling and text analysis. You work with an analytical mindset and try to answer questions as precisely as possible.

Prompt

Let us play a guessing game. Based on the given topic description, your task is to guess the user's <private feature>.
Topic Description:

<topic descriptions>

Output Format:

```
{
 "inference": "Your explanation here",
 "guess": "Your guess here",
 "confidence": "Your confidence score here, 0-1"
}
```

Only respond in the specified JSON format. Do not include any additional text.
Choose from these options: <options>

Model	Feature	Accuracy	E-AOL	E-ACB/CI (All samples)	E-ACB/CI (Only correct samples)	Total
Llama-3.3-70B-Instruct	Income	0.568	-0.573 (-0.869, -0.277)	-0.093 (-0.155, -0.032)	-0.103 (-0.165, -0.041)	412
	Gender	0.887	0.542 0.318, 0.766	0.247 (0.153, 0.341)	0.294 (0.202, 0.385)	274
	Age	0.665	-0.179 (-0.409, 0.050)	0.052 (-0.004, 0.108)	0.054 (-0.003, 0.112)	412
	Relationship	0.735	0.134 (-0.245, 0.513)	0.315 (0.127, 0.503)	0.492 (0.159, 0.824)	321

Table 3: Accuracy, E-AOL, and E-ACB corresponding to all data samples by an adversary before sanitization.

Model	Feature	After Sanitization Prompt		After Deception Prompt		Total
		E-AOL/ CI/ P-VAL	E-ACB/ CI/ P-VAL	E-AOL/ CI/ P-VAL	E-ACB/ CI/ P-VAL	
Llama-3.3-70B-Instruct	Income	-0.999 (-1.281, -0.717) 0.006	-0.184 (-0.246, -0.123) 0.001	-1.892 (-2.500, -1.285) 0.002	-0.049 (-0.121, 0.023) 0.995	412
	Gender	0.091 (-0.144, 0.326) 0.004	0.014 (-0.060, 0.087) 0.001	-0.887 (-1.296, -0.479) 0.001	0.230 (0.154, 0.307) 0.385	274
	Age	-0.328 (-0.568, -0.087) 0.006	-0.016 (-0.074, 0.043) 0.000	-1.242 (-1.628, -0.856) 0.000	0.064 (0.006, 0.121) 0.945	412
	Relationship	-0.440 (-0.739, -0.142) 0.009	0.175 (-0.011, 0.361) 0.004	-2.189 (-2.883, -1.496) 0.016	0.336 (0.177, 0.496) 0.385	321
Deepseek-R1-32b	Income	-1.204 (-1.647, -0.760) 0.006	-0.145 (-0.213, -0.077) 0.004	-1.318 (-1.667, -0.969) 0.006	-0.121 (-0.176, -0.066) 0.019**	412
	Gender	0.021 (-0.219, 0.260) 0.004	-0.001 (-0.078, 0.076) 0.001	-0.376 (-0.788, 0.037) 0.001	0.095 (0.025, 0.166) 0.001**	274
	Age	-0.405 (-0.697, -0.113) 0.004	-0.016 (-0.079, 0.048) 0.000	-0.358 (-0.658, -0.058) 0.014	0.031 (-0.026, 0.087) 0.003**	412
	Relationship	-0.444 (-0.757, -0.131) 0.018	0.193 (0.033, 0.353) 0.007	-0.896 (-1.561, -0.231) 0.014	0.270 (0.110, 0.430) 0.087	321
Gpt-4.1	Income	-0.977 (-1.397, -0.557) 0.037	-0.185 (-0.267, -0.104) 0.004	-2.078 (-2.732, -1.424) 0.008	-0.062 (-0.123, -0.001) 0.982	412
	Gender	-0.128 (-0.471, 0.214) 0.004	0.002 (-0.076, 0.080) 0.001	-1.711 (-2.472, -0.950) 0.008	0.190 (0.069, 0.312) 0.069	274
	Age	-0.509 (-0.763, -0.254) 0.003	-0.039 (-0.101, 0.023) 0.000	-0.860 (-1.370, -0.350) 0.003	0.052 (-0.013, 0.118) 0.384	412
	Relationship	-0.387 (-0.733, -0.040) 0.046	0.189 (0.018, 0.360) 0.001	-2.255 (-3.593, -0.918) 0.009	0.330 (0.155, 0.505) 0.384	321

Table 4: E-AOL and E-ACB values per feature after sanitization/deception for different models for all data samples

Model	Feature	After Sanitization Prompt		After Deception Prompt		Total Correct Pre-sanitization
		E-ACB/CI/P-VAL (correct)	E-ACB/CI/P-VAL (incorrect)	E-ACB/CI/P-VAL (correct)	E-ACB/CI/P-VAL (incorrect)	
Llama-3.3-70B-Instruct	Income	-0.189 (-0.243, -0.135) 0.002	-0.238 (-0.380, -0.095) 0.010	-0.113 (-0.195, -0.030) 0.527	-0.094 (-0.190, 0.002) 0.674	234
	Gender	0.016 (-0.055, 0.087) 0.001	0.013 (-0.072, 0.097) 0.002	0.253 (0.169, 0.338) 0.500	0.225 (0.105, 0.346) 0.080	243
	Age	-0.009 (-0.071, 0.053) 0.000	-0.046 (-0.123, 0.031) 0.001	0.065 (0.005, 0.125) 0.861	0.065 (-0.003, 0.132) 0.913	274
	Relationship	0.383 (0.024, 0.741) 0.042	0.189 (-0.104, 0.482) 0.008	0.291 (0.062, 0.520) 0.156	0.513 (0.114, 0.912) 0.539	236
Deepseek-R1-32b	Income	-0.133 (-0.182, -0.084) 0.213	-0.186 (-0.284, -0.087) 0.004	-0.114 (-0.171, -0.057) 0.248	-0.156 (-0.248, -0.064) 0.074	234
	Gender	0.023 (-0.063, 0.109) 0.001	-0.037 (-0.115, 0.040) 0.002	0.047 (-0.022, 0.115) 0.001	0.148 (0.045, 0.251) 0.024**	243
	Age	-0.010 (-0.075, 0.055) 0.000	-0.040 (-0.109, 0.028) 0.001	0.040 (-0.017, 0.097) 0.062	0.009 (-0.080, 0.097) 0.020**	274
	Relationship	0.356 (0.003, 0.709) 0.024	0.199 (-0.024, 0.421) 0.004	0.172 (0.050, 0.294) 0.037	0.493 (0.031, 0.956) 0.064	236
GPT-4.1	Income	-0.203 (-0.271, -0.136) 0.002	-0.202 (-0.304, -0.101) 0.004	-0.078 (-0.152, -0.004) 0.531	-0.107 (-0.195, -0.020) 0.633	234
	Gender	0.034 (-0.043, 0.112) 0.002	-0.021 (-0.121, 0.079) 0.002	0.057 (-0.107, 0.222) 0.008	0.207 (0.081, 0.334) 0.423	243
	Age	-0.038 (-0.101, 0.024) 0.000	-0.056 (-0.136, 0.024) 0.001	0.045 (-0.025, 0.116) 0.319	0.043 (-0.015, 0.101) 0.577	274
	Relationship	0.381 (0.004, 0.759) 0.002	0.153 (-0.072, 0.378) 0.002	0.049 (-0.184, 0.282) 0.022	0.538 (0.137, 0.938) 0.652	236

Table 5: E-AOL and E-ACB computed after sanitization and deception prompts, only to the subset of samples that were correctly inferred by the adversary before sanitization. Confidence intervals and corresponding p-values are also reported.

	Input token	**Output token**	**Total**
Sanitization technique (Our self-verification iterative)	494	393	887
Sanitization technique in [39]	3210	1155	4366

Table 6: Token count (in thousands) comparison between our proposed sanitization prompt mechanism and that of [39].

Model	Feature	After Sanitization Prompt		Total
		E-AOL/ CI/ P-VAL	E-ACB/ CI/ P-VAL	
Llama-3.3-70B-Instruct	Income	-1.105 (-1.292, -0.789) 0.005	-0.214 (-0.267, -0.145) 0.001	412
	Gender	0.071 (-0.244, 0.289) 0.005	0.002 (-0.080, 0.057) 0.001	274
	Age	-0.301 (-0.489, -0.067) 0.004	-0.011 (-0.064, 0.049) 0.000	412
	Relationship	-0.420 (-0.689, -0.122) 0.006	0.181 (-0.009, 0.300) 0.004	321
Deepseek-R1-32b	Income	-1.250 (-1.747, -0.848) 0.005	-0.162 (-0.223, -0.097) 0.003	412
	Gender	0.017 (-0.213, 0.220) 0.002	-0.002 (-0.065, 0.067) 0.001	274
	Age	-0.425 (-0.612, -0.118) 0.003	-0.011 (-0.066, 0.056) 0.001	412
	Relationship	-0.350 (-0.647, -0.091) 0.012	0.201 (0.034, 0.355) 0.005	321
Gpt-4.1	Income	-1.245 (-1.596, -0.894) 0.009	-0.221 (-0.299, -0.143) 0.004	412
	Gender	-0.336 (-0.672, -0.000) 0.001	-0.044 (-0.131, 0.044) 0.001	274
	Age	-0.429 (-0.691, -0.167) 0.001	-0.054 (-0.115, 0.008) 0.001	412
	Relationship	-0.602 (-1.153, -0.051) 0.006	0.175 (0.030, 0.320) 0.007	321

Table 7: E-AOL and E-ACB values per feature after sanitization using iterative prompt technique as proposed in [39]

Model	Feature	Sanitization			
		LLM Utility	Cosine Sim	BLEU	ROUGE
Llama-3.3-70B-Instruct	Income	0.909 (0.901, 0.914)	0.868 (0.861, 0.881)	0.371 (0.351, 0.399)	0.613 (0.599, 0.632)
	Gender	0.941 (0.934, 0.950)	0.931 (0.928, 0.938)	0.489 (0.450, 0.512)	0.632 (0.601, 0.665)
	Age	0.929 (0.922, 0.934)	0.901 (0.892, 0.909)	0.440 (0.423, 0.452)	0.662 (0.642, 0.676)
	Relationship	0.918 (0.914, 0.928)	0.917 (0.909, 0.927)	0.481 (0.471, 0.509)	0.731 (0.712, 0.749)
Deepseek-R1-32b	Income	0.912 (0.903, 0.917)	0.885 (0.877, 0.893)	0.329 (0.307, 0.335)	0.589 (0.566, 0.611)
	Gender	0.948 (0.946, 0.952)	0.929 (0.923, 0.934)	0.524 (0.501, 0.546)	0.630 (0.621, 0.660)
	Age	0.941 (0.938, 0.946)	0.891 (0.882, 0.898)	0.194 (0.129, 0.218)	0.505 (0.409, 0.549)
	Relationship	0.932 (0.924, 0.935)	0.901 (0.887, 0.909)	0.412 (0.387, 0.439)	0.666 (0.632, 0.678)
GPT-4.1	Income	0.894 (0.886, 0.903)	0.832 (0.824, 0.841)	0.241 (0.220, 0.267)	0.465 (0.445, 0.485)
	Gender	0.945 (0.939, 0.950)	0.883 (0.876, 0.890)	0.258 (0.235, 0.280)	0.521 (0.498, 0.544)
	Age	0.921 (0.915, 0.927)	0.829 (0.821, 0.838)	0.134 (0.120, 0.148)	0.350 (0.333, 0.367)
	Relationship	0.937 (0.931, 0.944)	0.907 (0.900, 0.914)	0.369 (0.346, 0.392)	0.625 (0.604, 0.646)

Table 8: Comparison of GPT, LLaMA3, and DeepSeek across features and metrics after sanitization using iterative prompt technique as proposed in [39] with 95% confidence intervals.

Feature	Original text	Sanitized text	BLEU	ROUGE	Cosine Similarity
income_level	Strict meal prepping; definitely stretches those pounds! .Time for home-cooked gourmet then!	Being mindful about meals; definitely helps!.Time to prepare food then!	0.39	0.33	0.78
income_level	Ditched car altogether; biking works wonders—saves cash & keeps ya fit! .DIY projects at home doubled as hobby time for me!	Opted for a different way to get around—biking is both healthy and cost-effective..I enjoy spending time on various hobbies and projects at home.	0.35	0.3	0.79
income_level	Mastered quality meals with less $$!	Found ways to prepare delicious meals efficiently!	0.18	0.17	0.8
age	finding those secret spots sounds ace but when your schedule's back-to-back work & lectures discover time's like hunting unicorns	discovering those hidden spots sounds great but when your schedule's packed with responsibilities, finding time feels impossible	0.57	0.45	0.82
age	man those quirky stationary shops downtown? they're swanky boutiques now - kinda miss digging for unique postcards among shelves piled high with notebooks... .noticed some classic diners turning into fancy pet cafes around here	shops have become boutiques - miss browsing for postcards and notebooks....some places have become pet cafes	0.22	0.36	0.82

Table 9: Pairs of original and sanitized texts illustrating low lexical overlap (BLEU/ROUGE) vs. high semantic similarity (cosine).