Truthful Text Sanitization Guided by Inference Attacks

Ildikó Pilán[a], Benet Manzanares-Salor[1b], David Sánchez[b], Pierre Lison[a]

[a] *Norwegian Computing Center, Postboks 114 Blindern, Oslo, NO-0314, Norway*
[b] *Department of Computer Engineering and Mathematics. CYBERCAT, Universitat Rovira i Virgili, Tarragona, 43007, Spain*

Abstract

Text sanitization aims to rewrite parts of a document to prevent disclosure of personal information. The central challenge of text sanitization is to strike a balance between privacy protection (avoiding the leakage of personal information) and utility preservation (retaining as much as possible of the document's original content). To this end, we introduce a novel text sanitization method based on *generalizations*, that is, broader but still informative terms that subsume the semantic content of the original text spans. The approach relies on the use of instruction-tuned large language models (LLMs) and is divided into two stages. Given a document including text spans expressing personally identifiable information (PII), the LLM is first applied to obtain truth-preserving replacement candidates for each text span and rank those according to their abstraction level. Those candidates are then evaluated for their ability to protect privacy by conducting *inference attacks* with the LLM. Finally, the system selects the most informative replacement candidate shown to be resistant to those attacks. This two-stage process produces replacements that effectively balance privacy and utility.

We also present novel metrics to evaluate these two aspects without needing to manually annotate documents. Results on the Text Anonymization Benchmark show that the proposed approach, implemented with Mistral 7B Instruct, leads to enhanced utility, with only a marginal (< 1 p.p.) increase in re-identification risk compared to fully suppressing the original spans. Furthermore, our approach is shown to be more truth-preserving than existing methods such as Microsoft Presidio's synthetic replacements.

Keywords: data privacy, text sanitization, data utility, truth-preserving

[1]Corresponding author. Email: benet.manzanares@urv.cat

1. Introduction

Text documents play a key role in many sectors, including healthcare, legal services and public administrations. However, those documents often contain personal data that, if released without adequate protection, can pose significant privacy risks to the individuals they concern. Regulations such as the General Data Protection Regulation (GDPR) [24] impose strict constraints on how personal information should be protected. In particular, personal information cannot be shared with third parties or employed for secondary purposes without a proper legal ground, such as the explicit consent of the individuals to whom the data refers.

One strategy to overcome those constraints is to *anonymize* the data, in which case the data is no longer considered personal, and thus outside the scope of those regulations [24]. Beyond regulatory compliance, anonymization also helps balance the pursuit of data-driven research and innovation with the protection of individual rights. By reducing the risk of reidentification, anonymization fosters greater trust between individuals and institutions, encouraging more open data sharing practices that ultimately benefit society as a whole.

Multiple text anonymization techniques have been proposed in the literature [36]. In particular, *text sanitization* prevents the disclosure of personal information by: (i) *detecting* spans that contain personally identifiable information (PII), and (ii) *masking* those spans. Most works have focused on the first step (PII detection), and very often rely on simple masking techniques, such as removing PIIs or replacing them by generic semantic categories such as PERSON or LOCATION) [32, 64, 65, 29, 37, 10, 44]. However, this relatively crude level of protection can undermine the downstream utility of the sanitized texts, which is the ultimate goal of data sharing. As the data privacy literature has highlighted [30], data protection should aim at balancing the trade-off between the attained level of privacy and the preservation of data utility. For textual data, this means designing effective *text sanitization* mechanisms that preserve the original text's meaning as much as possible while satisfying the privacy needs of the data release.

In this context, large language models (LLMs), which have shown strong capabilities in a wide range of natural language understanding and generation tasks, have emerged as promising tools for enhancing text sanitization.

Recent studies have demonstrated their effectiveness in text editing [58] and simplification [33] with strong zero- and few-shot performance due to instruction tuning [42].

Despite these advances, current text sanitization approaches rarely leverage LLMs to *generate* replacements for detected PII spans. Most existing methods rely on suppression or named entity-based substitutions, which fail to produce coherent and realistic text. As a result, there is currently no principled method for generating *truthful and utility-preserving replacements* of sensitive text spans. Addressing this gap is essential for enabling safe and effective data sharing under privacy regulations while maintaining the semantic integrity and analytical value of textual data.

1.1. Contributions and plan

In this paper, we propose a novel text sanitization approach, named INference-guided Truthful sAnitization for Clear Text (INTACT). INTACT leverages instruction-tuned open-weights LLMs to replace personal information with less specific, but truth-preserving *generalizations* that retain the original semantic content to the greatest extent possible, while preventing inference attacks from recovering the initial data.

Our work brings the following contributions:

- Unlike previous privacy-oriented text rewriting methods [6, 18], the presented method adopts a two-stage strategy. Given a text span to replace, we first (1) generate a sorted list of generalization candidates at different abstraction levels, and then (2) select the most suitable replacement among those alternatives. This selection of the best replacement is done by testing whether the LLM can guess the original span from the replacement and the rest of the text (*inference attacks*). LLMs have indeed shown to be effective adversaries in anonymization tasks in previous studies, either on their own [45] or augmented with a retrieval component [12]. The combination of those two steps not only allows for balancing and customizing the privacy-utility trade-off, but also enables a more granular evaluation of the rewriting process.

- We present an evaluation framework for text sanitization that measures privacy and utility without reference annotations. This framework includes a novel utility metric, Text Preserved Similarity (TPS), which measures the similarity between original and sanitized documents according to both information content and semantic similarity.

- We evaluate and compare our method against several baselines on the Text Anonymization Benchmark (TAB) [46], a dataset explicitly tailored for the evaluation of text sanitization methods. Empirical results show that our method achieves the best privacy-utility trade-off, while keeping the protected data truthful to their original content.

As text sanitization is by definition a process that involves the processing of (sometimes highly sensitive) personal data, our approach relies on the use of open-weights LLMs, as those can be run locally without necessitating the transfer of data to third-party cloud solutions.

The remainder of the paper is structured as follows. Section 2 provides an overview of the text anonymization task, and discusses related works on text sanitization and text anonymization evaluation. Section 3 presents our text sanitization method, INTACT, and Section 4 describes our utility and privacy evaluation metrics. Section 5 gives implementation details and reports results and evaluation figures. Section 7 provides the conclusions and outlines future research directions.

2. Background and related works

2.1. Problem overview

Text anonymization aims to remove or mask personally identifiable information (PII) that may lead to re-identifying the subject(s) referenced in the text and/or disclose their confidential information. This process should consider all PII, including both *directly identifying* information –that is, information that refers to single subjects, such as full names or passport numbers–, and *quasi-identifying* terms which do not enable re-identification when considered in isolation, but may do so when combined with other quasi-identifiers, such as the combination of gender, birth date and postal code.

For detecting PII, the majority of automatic text anonymization approaches rely on *Named Entity Recognition* (NER) [32, 64, 65, 29, 37, 10]. Indeed, due to their specificity, named entities (NEs) often increase the re-identification risk. However, because not all PII fit within the limited set of NE types supported by NER tools, NER-based methods may miss a significant amount of disclosive terms [36]. To mitigate this issue, other detection approaches grounded on the broader notion of disclosure risk have been proposed [27, 44, 51]. In parallel, some methods based on differential privacy [19] have been proposed for text protection [22]. These typically treat texts

as bags-of-words, replacing sensitive terms via the exponential mechanism or adding Laplace noise to term counts. However, they often degrade utility and readability, failing to preserve truthfulness and document structure.

Once disclosive terms are detected in a first processing stage, two editing strategies are possible. Whereas document *redaction* directly suppresses or blacks out terms, thereby maximizing privacy at the cost of the utility and readability of the protected documents, text *sanitization* replaces PII spans with broader, more abstract terms [9]. By obscuring the disclosive information while preserving as much as possible of the original text's semantic content and readability, text sanitization seeks to offer a better privacy-utility trade-off than redaction. The choice of replacement strategy plays a crucial role in this balance. Generalizations, which abstract sensitive terms without distorting their meaning, help maintain the truthfulness of the document and are particularly valuable in domains like medicine or law, where accuracy is essential. In contrast, substituting terms with unrelated or misleading ones (*e.g.*, replacing broken arm with bird flu) can compromise the integrity of the information and lead to serious misinterpretations.

2.2. Sanitization

The majority of text anonymization methods use either strict redaction or replace the disclosive terms by the label of their corresponding category, which is a basic form of sanitization [36]. Both strategies retain little to no meaningful information from the original terms. More sophisticated approaches leverage structured knowledge bases, such as ontologies, or, more recently, language models, to retrieve potential replacements. We survey those methods below.

2.2.1. NER-based sanitization

Following the broad usage of NER for PII detection, the most common practice for replacements is to use the named entity labels (*e.g.*, PERSON or LOCATION). This straightforward approach maintains truthfulness to the extent that the classification is accurate, because the labels themselves offer a coarse generalization of the original terms. A notable exception is the MISC (miscellaneous) category [61], which is used in some NER systems to cover entities that do not fit into more specific categories. Replacing such entities with MISC is effectively equivalent to redaction, as it conveys no semantic information about the original term. Overall, NER-based sanitization offers

limited utility preservation, as the original content of the PII spans is mostly lost and the readability of the text is severely reduced.

2.2.2. Sanitization from knowledge bases

Other text sanitization methods have leveraged structured knowledge bases [14, 4, 3, 53, 54, 52, 27, 43], such as WordNet, the now defunct Open Directory Project (ODP) or YAGO, to retrieve term hypernyms (*i.e.*, generalizations) to be used as replacements. Those methods rely on dedicated mechanisms to select the most appropriate generalization from a list of candidates. In [27], the authors leverage the Word2Vec [41] word embedding model to identify terms that are semantically related to the individual to be protected. If a term's embedding has a cosine similarity with the individual's name above a user-defined threshold, it is considered disclosive, and it is replaced by its most specific generalization (retrieved from YAGO) whose cosine similarity falls below the threshold. In [53, 52] the authors employ a similar approach, but this time selecting the most appropriate replacement through measures of the term informativeness and mutual information. [4] propose to select generalizations so that they encompass a number of specializations (or hyponyms) that would result in t or more *plausible* documents. [3] employs a contingency table that quantifies the degree of correlation between each pair of textual terms and a taxonomy modeling term generalizations. Finally, [43] rely on human judgments to select the most appropriate generalization for each sensitive term based on its context.

Nonetheless, sanitization methods based on "static" knowledge bases have limited applicability due to their relatively narrow coverage of disclosive terms. This limited coverage is particularly salient for named entities, which are rarely present in such knowledge bases, while they are often one of the most crucial type of terms to sanitize. Moreover, the largest knowledge bases, such as YAGO or ODP may not provide consistent generalizations, as they originate from arbitrary categories, labels and descriptions (such as Wikipedia categories) that do not systematically satisfy the *subsumption property* – that is, the requirement that a generalized term should denote a concept that logically encompasses all instances of the original text span.

2.2.3. Sanitization with language models

A number of recent studies have also explored the use of language models (LMs) for text anonymization, both for detecting and for protecting sensitive text spans. Such models successfully overcome the aforementioned coverage

issues of knowledge bases, as they can (1) generate replacements for any input span and (2) ensure those replacements integrate seamlessly into the surrounding text, preserving grammaticality and fluency.

The MAPA anonymization toolkit [6] is an open source tool that supports the detection and replacement of a variety of PII types. The replacements are sampled from the most likely tokens based on BERT language models fine-tuned for a variety of European languages and domains including administration, health and law. [2] rely on similar models, but without fine-tuning for PII replacement. In both cases, the tokens of the sensitive terms are initially replaced by the special token `[MASK]`. The BERT model then predicts for each mask a distribution over possible tokens based on the context. The replacements are finally sampled from these distributions. Due to the use of small-size LMs, those approaches only require modest computational resources, but exhibit several limitations. First, the generated replacements are constrained to match the token length of the original sensitive span, preventing substitutions that are shorter or longer. This constraint not only limits flexibility but also introduces a vulnerability, as adversaries may exploit this token-length consistency to infer the original content. Moreover, there is no mechanism to ensure that the replacement remains truthful to the original PII span.

Larger, instruction fine-tuned LLMs such as ChatGPT [63] have also been proposed for protecting sensitive information. Instead of fine-tuning, a common way to get these models to perform a task is via an instruction (*prompt*) and optional examples. Such prompting methods are also integrated in existing de-identification tools, such as Microsoft Presidio [39]. Given a text, the PII detected with the Presidio pipeline can be replaced with GPT 3.5-generated alternatives in a single step[2]. Concretely, the LLMs receives the document to be processed with the sensitive spans already replaced by their NE type (*e.g.*, PERSON or LOCATION), and is prompted to generate replacements of the same type. This makes it possible to use a commercial, cloud-based LLM without having to send the original document to third parties. Nevertheless, this method requires complete confidence in the detection step, with major privacy implications if any PII is missed. Furthermore, like for the aforementioned BERT-based approaches, truthfulness is neglected, with

[2]See `https://github.com/microsoft/presidio/blob/main/docs/samples/python/synth_data_with_openai.ipynb`

the generated replacements only restricted to be of the same NE type as the original span.

The only LLM-based approach explicitly focused on generating *truthful* replacements (that is, replacements that subsume the semantic content of the original text span) is the one described in [18]. This study targets the task closely related to sanitization of abstracting text spans containing *self-disclosure*, "the communication of personal information to others". The categories of personal information employed in this study are a fine-grained version of those typically considered by previous anonymization approaches, and mostly fall under demographic attributes. The sensitive text spans in [18], however, are often phrases and clauses (written in first person) rather than words or named entities. Their proposed task can be thus viewed as a specialized case of protecting sensitive information. PII protection is performed step-wise, where self-disclosure spans are first detected with a BERT model fine-tuned on a manually annotated dataset. The text spans expressing personal information are then re-written with an open-weights instruct LLM (Llama-2-7B) fine-tuned on a dataset with replacements automatically generated using GPT-3.5 and GPT-4.

In contrast, the work presented in this paper only requires access to an open-weights instruct LLM, without additional fine-tuning, to iteratively select *truthful* generalizations from a set of alternatives with different abstraction levels. In addition, the proposed method performs inference attacks to test whether a given replacement effectively conceals the initial content of the span.

Table 1 summarizes key characteristics of representative LLM-based sanitization methods, highlighting the underlying models, their methodologies, whether they ensure truthfulness in the replacements, and any notable limitations or contributions.

2.3. Evaluation

A common evaluation strategy for text anonymization methods is to compare their editing choices against manual annotations, which are considered as ground truths. The anonymization quality is quantified via precision and recall metrics, which are used as proxies of the preserved utility and the attained privacy protection, respectively [36, 51, 26, 32]. Modified versions of the precision and recall metrics specifically tailored for text anonymization have also been proposed [46].

Method	Approach	Truthful?	Model	Other details
MAPA toolkit [6] and Albanese (2023) [2]	Masking and sampling from top token predictions	✗	BERT (with possible fine-tuning)	Restricted to same-length replacements. Vulnerable to inference attacks
Presidio's synthetic replacements [39]	Prompting with sensitive spans replaced by NE types	✗	GPT-3.5-turbo	Uses closed-weights LLM. No semantic preservation guarantees
Dou et al. (2024) [18]	Fine-tuning and prompting for abstraction	✓	Fine-tuned LLaMA-2-7B	Targets self-disclosure. Training data partially based on GPT-3.5/4. Uses open-weights LLM
INTACT (this work)	Truth-preserving replacements + inference attacks	✓	Mistral-7B-Instruct-v0.2	No fine-tuning. Evaluation of privacy risks via inference attacks. Uses open-weights LLM

Table 1: Comparison of language model-based sanitization methods.

Precision measures the percentage of text spans replaced by the automatic method that human annotators also deemed disclosive. Because any additional replaced text spans are deemed unnecessary, lower precision figures are interpreted as poorer utility preservation. On the other hand, recall quantifies the percentage of disclosive text spans, as identified by human annotators, that the automatic method replaced. Because untouched disclosive spans are considered privacy failures, lower recall figures are understood as weaker privacy protection.

However, these evaluation metrics are only applicable to assess the quality of the detection step, and are unable to assess the suitability of the replacements used for sanitization. Indeed, evaluating sanitized texts against a manually constructed ground truth is challenging because i) there can be multiple valid (*i.e.*, non-disclosive) combinations of replacements for a given text and, ii) replacement candidates are, quite often, unbounded. Establishing a single ground truth for the evaluation is thus challenging. Moreover, requesting hu-

man annotators to provide privacy-preserving replacements (or, at the very least, chose appropriate replacements for predefined lists of candidates) requires even more effort than the –already quite costly– detection of disclosive text spans [43]. As a result, the sanitization strategies currently proposed in the literature remain largely untested [14, 31, 3, 53, 54, 52, 27, 43].

3. Text sanitization method

Our text sanitization approach, INTACT, is based on instruction-fine-tuned LLMs and breaks down the task into two separate steps:

- Generating replacement candidates for disclosive terms with different levels of generalization
- Selecting the most appropriate replacement among those candidates through the use of inference attacks.

The purpose of this two-stage process is to select replacements that satisfy three conditions:

1. they should be *privacy-preserving*, i.e. they should conceal the original content of the span, also to adversaries seeking to infer their values based on the selected replacements and the context they appear in.
2. they should be *truthful*, i.e. they should subsume the semantic content of the original span.
3. they should be *as informative as possible* (to maximize the data utility) given that the two above conditions are satisfied.

A high-level visual overview of INTACT is presented in Figure 1 together with an example replacement generation and selection outcome.

We will assume throughout this paper a document D in which we have detected a set of text spans $T = \{t_1, t_2, \ldots, t_n\}$ that may disclose personal information and ought to be edited. The task of detecting those PII-related spans in text documents is outside the scope of this paper, but has already been studied in numerous works, such as [40, 36, 46]. Common approaches to this task rely on standard NER tools or sequence labeling models tailored for detecting PII-related information. We will also assume that each of those detected text spans t_i in D has been tagged with a corresponding entity type label l_i, such as ORG or PERSON.

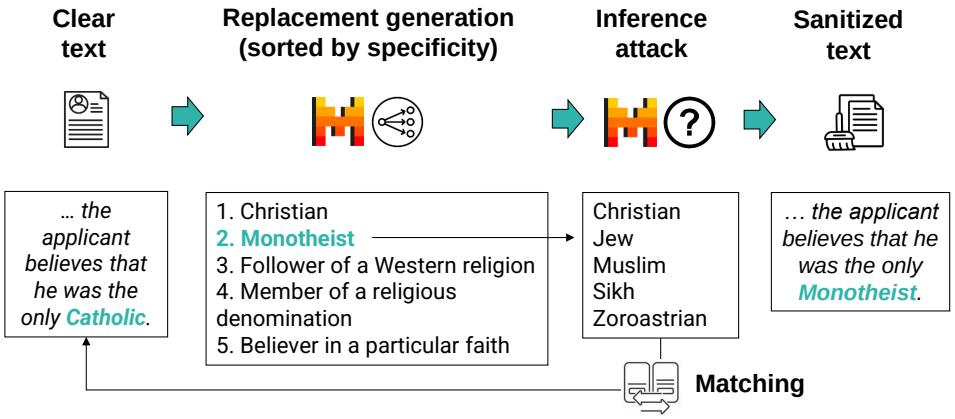


Figure 1: Illustration of text sanitization process with INTACT.

3.1. Replacement generation

In this first step, we prompt an LLM to generate a list of generalized replacement candidates for the provided text span in its context. We explicitly request replacements with different levels of abstraction that the LLM is then required to sort from most to least specific. This sorting is done within the same step as generation, in order to limit computational costs. To enhance the quality of the generated candidates, we also include a one-shot example containing a sentence with a target span (of the same label as the span to sanitize) and some possible generalizations. We provide the text span in its local context, defined here as the sentence it occurs in. Figure 1 shows an example with (partial) context and sorted replacements for *Catholic*, where *Christian* is listed as the most specific replacement and *Believer in a particular faith* as the least specific one.

Formally, given an instruction fine-tuned language model $\texttt{LLM}_{\texttt{instr}}$ and an input text span t_i occurring in D that should be sanitized, we prompt the LLM to generate a list of m replacements candidates $C_i = \{c_{i,1}, c_{i,2}, \ldots, c_{i,m}\}$, where the candidates $c_{i,j}$ are sorted by order of abstraction. Sorting is performed by the LLM during generation, based on the instructions given in the candidate generation prompt (provided in Figure 2), which request that candidates be sorted from most specific to most generic. The list of replacement candidates C_i for the text span t_i is thus defined as:

$$C_i = \texttt{LLM}_{\texttt{instr}}(p_C, e_C, t_i, s_i) \tag{1}$$

where p_C is the prompt used to instruct $\texttt{LLM}_{\texttt{instr}}$ to generate the candidates, e_C is the one-shot example, t_i is the input text span that should be edited, and s_i is the sentence in D containing that text span. Including sentence-level context is intended to provide sufficient information to ensure grammatical sanitized outputs and to address potentially ambiguous PII.

The text spans $\{t_1, t_2, \ldots, t_n\}$ are processed according to their order of occurrence in D. For entities that occur multiple times through the same document D, we only generate replacements for the first mention of that entity and reuse the selected replacement for all subsequent mentions.

For direct identifiers – such as person names or phone numbers – which can reveal identity even in isolation, replacement candidates are not obtained through Eq. 1. Instead, in line with how direct identifiers are handled in statistical disclosure control [30], direct identifiers are replaced by placeholders such as entity labels, for instance PERSON i. Those replacements, implemented through heuristic rules, not only help maintaining privacy, but it also prevents unintended identification, *i.e.*, names or codes matching actual entities. To preserve co-reference chains, the heuristics assign consistent labels to repeated mentions of the same entity. Rule-based generalisations are also employed to produce consistent generalizations for dates expressed in standard formats, as those can be generated directly from heuristics without relying on LLMs (and thus at a much lower computational cost). See Section 5.2 for more details on those generalization heuristics.

3.2. Replacement selection

The second step of the method identifies the *best* replacement c_i^* among the possible replacement candidates C_i for the original text span t_i. The best replacement is here defined as the most specific candidate that does not reveal, directly or indirectly, the original content of the span t_i. In other words, the best replacement should be the most informative replacement among those that do not constitute a disclosure risk. To determine whether a replacement candidate $c_{i,j}$ constitutes a disclosure risk, an *inference attack* is carried out – that is, we use the instruction-tuned LLM to produce a set of guesses for the original span given some context. The inference attacks are performed based on one-shot prompting, without additional fine-tuning of the LLM for this specific task. These guesses are then compared to the original span to assess whether the replacement is acceptable (*i.e.* if none of the guesses match the original span) or whether it would allow an adversary to find back the original text span. Crucially, those guesses are generated by

taking into account not only the replacement candidate, but the full (edited) document as context, in order to allow the LLM to take advantage of all information expressed in the document to predict the original content of a given span. For instance, in the example in Figure 1, the LLM cannot infer *Catholic* when it is replaced by *Monotheist* in the original document-long context, it is therefore deemed as a safe replacement.

Inference attack. For each text span t_i and replacement candidate $c_{i,j}$ for this span, we generate a set of p guesses $G_{i,j} = \{g^1_{i,j}, g^2_{i,j}, \ldots, g^p_{i,j}\}$ by prompting the instruction-tuned LLM:

$$G_{i,j} = \texttt{LLM}_{\texttt{instr}}(p_G, e_G, c_{i,j}, D') \tag{2}$$

in which:

- p_G is the one-shot prompt employed for the generation of guesses.
- e_G is an one-shot example composed of a short synthetic text, with one span substituted with a synthetic LLM replacement and a set of synthetic LLM guesses.
- $c_{i,j}$ is the candidate replacement for the original input text span t_i.
- D' is the current version of the sanitized document. More precisely, D' is obtained from the original document D by replacing each original text span t_j with its best replacement c^*_i, if it has already been selected, or with the most specific replacement candidate $c_{i,1}$ if is has not.

Matching. To evaluate the effectiveness of those attacks, it is necessary to define what constitutes a successful guess from the point of view of the adversary. A naive approach would be to define a guess as successful if it returns the exact same string as the original text span. This approach, however, fails to account for guesses that are closely related but not entirely identical to the original string. To capture those variations, we define a $\textsc{match}(t_i, g^k_{i,j})$ condition, which returns `true` if any of the following conditions hold:

- *Lemma Overlap*: There is an overlap between the set of lemmas (base forms of words, *e.g.*, *dog* for *dogs*) of t_i and $g^k_{i,j}$.
- *N-gram Overlap*: If t_i is a named entity, an overlap exists between the n-grams (sequences of n characters) of t_i and $g^k_{i,j}$.

Non-alphabetic tokens, stop words and high-frequency lemmas being sanitized are excluded from both t_i and $g^k_{i,j}$ when computing lemma overlap. Acronyms are handled by merging the first letters of all title-cased words in t_i into one word, doing the same for $g^k_{i,j}$, and expanding the respective lemma lists with these automatically coined (potential) acronyms. In all comparisons, the text spans are normalized to lowercase characters.

Based on this matching function, we can then define a replacement candidate $c_{i,j}$ as risky from a privacy perspective if at least one of the guesses matches the original text span t_i. More formally, we can thus define the following boolean function:

$$\textsc{risky_replace}(t_i, c_{i,j}) = \begin{cases} \mathsf{true} & \text{if } \exists\ g^k_{i,j} \in G_{i,j} \text{ where } \textsc{match}(t_i, g^k_{i,j}) = \mathsf{true} \\ \mathsf{false} & \text{otherwise} \end{cases}$$

3.3. Full sanitization process

The selection of the best replacement c^*_i among possible alternatives thus proceeds as follows:

- For a particular text span t_i to sanitize, we generate the replacement candidates C_i according to Eq. (1)
- For each replacement candidate $c_{i,j} \in C_{i,j}$, we generate the guesses on the content of the original text span based on Eq. (2)
- The candidates $c_{i,j} \in C_i$ are then examined in the order they were generated by the LLM. For each candidate, we compare its corresponding guesses to the original text span t_i.
- Finally, we select as replacement c^*_i the first replacement candidate $c_{i,j}$ for which RISKY_REPLACE returns false. If no candidate is found, as a fallback strategy, the text span is replaced by the name of its corresponding entity type l_i, such as LOCATION[3].

This process is detailed in Algorithm 1. For the implementation details, including the prompts used for replacement candidate generation and selection, see Section 5.2 and Appendix A.

[3]In the absence of entity labels, a placeholder such as "***" can be used.

Algorithm 1 Text sanitization procedure SANITIZE(D).

Input: An input text document D
Output: A sanitized document D'
Require: An instruction-tuned $\texttt{LLM}_{\texttt{instr}}$

 # Detect the disclosive text spans in D and their semantic categories
 # (the implementation of this function is outside the scope of this paper)
$T, L \leftarrow$ DETECT_DISCLOSIVE_SPANS(D)

 # Step 1: find replacement candidates for each disclosive span
for $t_i \in T$ **do**
  Extract sentence s_i in D in which t_i occurs
  # Generate sorted candidates with prompt p_C and one-shot e_C
  $C_i \leftarrow \texttt{LLM}_{\texttt{instr}}(p_C, e_C, t_i, s_i)$
end for

Initialize $D' \leftarrow D$ where all spans t_i (for $1 \leq i \leq n$) are initially replaced by their most specific replacement candidate $c_{i,1}$

 # Step 2: conduct inference attacks to determine best replacement
for $i \leftarrow 1$ to n **do**
  for $j \leftarrow 1$ to m **do**
    # Generate guesses about the original span based on candidate $c_{i,j}$
    $G_{i,j} = \texttt{LLM}_{\texttt{instr}}(p_G, e_G, c_{i,j}, D')$
  end for
  # Select candidate that is most specific and not risky
  $c_i^* \leftarrow \min(j \text{ with } 1 \leq j \leq m \text{ and } \text{RISKY_REPLACE}(t_i, c_{i,j}) = \mathsf{false})$
  # Fall back to a default value if all replacements were risky
  if $\text{RISKY_REPLACE}(t_i, c_{i,j})\ \forall\ j : 1 \leq j \leq m$ **then**
    $c_i^* \leftarrow$ semantic label l_i or placeholder
  end if
  Update D' by replacing t_i by c_i^*
end for

return D'

4. Evaluation metrics

As discussed in Section 2.3, current evaluation methods for text anonymization mostly focus on the task of *detecting* the occurrence of PII spans, while the subsequent problem of *replacing* those spans remains understudied and lacking appropriate evaluation metrics. To address this shortcoming, we propose two novel automatic evaluation metrics, respectively called TPS (Text Preserved Similarity) and TRIR (Text Re-identification Risk). Instead of relying on manual annotations, those metrics quantify *data utility* as a function of the proportion of semantic content preserved in the selected replacements relative to the original text span, and *privacy protection* as the inverse of the observed residual re-identification risk.

4.1. Utility assessment

The utility purpose of anonymized data is to produce outcomes in downstream tasks that closely align with those resulting from the original data.

In SDC for structured databases [30], this alignment is evaluated by measuring the *similarity* between the anonymized and the original data, as the two are known to be correlated. However, while quantifying similarity for structured numerical data is straightforward, doing so for unstructured text data is more challenging.

We propose a new utility metric for text anonymization that quantifies the similarity between an original document and its sanitized version from a semantic perspective, as the preservation of the semantic content of the text is directly related to its utility. Our proposal assesses the impact on utility of these replacements based on two factors:

1. the *information content* of the disclosive text span to be protected within the original document, which accounts for the semantic content encompassed by the original text span.
2. the *semantic similarity* of the replacement relative to the original text span, which expresses the proportion of semantic content preserved by the replacement.

4.1.1. Information content

In natural language processing, the informativeness of a text span, or its *information content* (IC), can be expressed according to Shannon's information theory as the negative logarithm of its probability of occurrence

[57]:

$$\mathrm{IC}(t) = -log(\Pr(t)) \tag{3}$$

This implies that the less probable (or predictable) a text span t is, the more information it provides. Conversely, if the text span is highly probable (or predictable), its information content is low.

Under this premise, we calculate the information content of a document D (that is, the amount of semantics it conveys) by summing the IC values of all its text spans. We name this *text information content* (TIC):

$$\mathrm{TIC}(D) = \sum_{t \in D} \mathrm{IC}(t) \tag{4}$$

As all text spans we consider the set of non-overlapping segments that are either a noun phrase, a named entity or a word other than a stop word.

We then measure the *relative information content* (RIC) of a text span t within a document D as the proportion of the text span's IC relatively to the TIC of D:

$$\mathrm{RIC}(t, D) = \frac{\mathrm{IC}(t)}{\mathrm{TIC}(D)} \tag{5}$$

Various methods are possible to practical compute the probability $\Pr(t)$ in Eq. 3, such as using their frequency of appearance within a corpus [49] (even leveraging the Web as corpus [53, 51]), considering the number of hypernyms and/or hyponyms within an ontology [56, 8], or evaluating their predictability with a language model [46]. The latter approach utilizes the general language understanding capabilities of language models (specifically BERT [16]) to predict words based on context. The model receives the context of the text span (*e.g.*, the surrounding sentence) with the target text span substituted by the [MASK] token. The probability of the text span is then determined by the model's probability of predicting it from the context. This LM-based method makes it possible to estimate the information content of any PII span without the need to rely on an in-domain corpus or ontology. For multi-token text spans, the minimum probability is selected, thereby ensuring that the measure of information content reflects the least predictable (and thus most informative) token.

As in [46], we rely on BERT (bert-base-uncased[4]) for computing the IC of our text spans. However, contrary to [46], we must in this case compute

[4] `https://huggingface.co/google-bert/bert-base-uncased`

the IC for all text spans within a document. Using the method proposed in [46] would involve substituting all text spans with [MASK] simultaneously, depriving the model of the necessary context to make accurate probability predictions. To solve this issue, we alternate the [MASK] substitution across text spans. Specifically, placing [MASK] for one of every N text spans, using the remaining spans as context. This technique requires making multiple predictions, each time shifting the [MASK] placements. For instance, with $N = 2$, two predictions are needed: one with [MASK] applied to the odd text spans while using the even spans as context, and another with [MASK] applied to the even spans while using the odd spans as context. Increasing N linearly increments the number of predictions (and thus the computational cost), but also helps refining the probability estimates by providing more context for each prediction.

4.1.2. Semantic similarity

As mentioned above, the *semantic similarity* between an original text span and its replacement indicates how well the semantic content of the former is preserved by the latter. A well-established measure of semantic similarity between text spans consists of calculating the cosine similarity between their embeddings derived from a language model [62, 15, 5].

A cosine similarity of 1 corresponds to identical spans, while 0 indicates no particular similarity, and -1 depicts completely unrelated text spans. Given that replacements are related to the original text span, we constrain the similarity calculation to the $[0, 1]$ range:

$$\textsc{Similarity}(t_i, c_i^*) = \max(0, \textsc{cos_sim}(\textsc{embed}(t_i), \textsc{embed}(c_i^*))) \tag{6}$$

where t_i is the original text span, c_i^* the selected replacement span, and EMBED is the embedding model.

As embedding model, we leverage *paraphrase-albert-base-v2*[5] in our experiments, an off-the-shelf Sentence-BERT model [48] trained in paraphrasing data, which has shown strong results on semantic similarity tasks [48, 35]. Given the proximity between paraphrasing and generalization, this model is well-suited to capture the semantic shifts resulting from text sanitization.

[5] `https://huggingface.co/sentence-transformers/paraphrase-albert-base-v2`

4.1.3. Text preserved similarity

Building upon the aforementioned measures, we now introduce *text preserved similarity* (TPS), an utility metric for text sanitization:

$$\text{TPS}(D, D') = \sum_{(t_i, c_i^*) \in (D, D')} \text{RIC}(t_i, D) * \textsc{Similarity}(t_i, c_i^*) \tag{7}$$

Where D' represents the sanitized document, D the original document, t_i the original span and c_i^* its selected replacement.

For unedited text spans, c_i^* is identical to t_i, resulting in a similarity of 1 and thus no loss in utility. TPS can be interpreted as the arithmetic mean of semantic similarities for all text spans, weighted by their relative informativeness. The TPS metric ranges from 0, indicating a loss of all original content, to 1, where D' conveys precisely the same meaning as D (*e.g.*, no text spans have been altered).

4.2. Privacy assessment

The ultimate goal of data anonymization is to minimize the re-identification risk. Considering that the greatest disclosure risk comes from the *combination* of disclosive text spans [36], a sound privacy evaluation should consider the residual re-identification risk of the chosen replacements as a whole. In structured data, this is done through record linkage attacks, which are the standard for empirical risk assessment [1, 17]. Building on this concept, [38] introduced an automatic metric called *text re-identification risk* (TRIR), which adapts the principles of record linkage to text, providing an objective way to assess the residual re-identification risk of anonymized documents.

More specifically, TRIR involves simulating a re-identification attack on the protected documents by leveraging, as background knowledge, a comprehensive collection of public documents of a population of individuals that encompasses the actual individuals to be protected. For example, if the individuals to be protected are hospital patients and the anonymized documents are their medical reports, all the citizens of the hospital's town could be the population, and the public documents leveraged as background knowledge could be a collection social media posts of those citizens.

To construct the attack, a language model with a classifier head is first trained to exploit this background knowledge to predict the individual corresponding a given document. This classification model is thus trained to identify the information most closely linked to each individual from the public documents. When presented with an anonymized document, the model

can then leverage the learned knowledge to re-identify the corresponding individual. If the predicted public individual is in fact the corresponding individual to be protected, re-identification is successful. TRIR is computed as the re-identification accuracy on the set of anonymized documents, which is in line to how record linkage is quantified for tabular data. Importantly, TRIR does not assume that the protected document is an abridged version of an original document, but any text containing information about an individual. This ensures full compatibility with text sanitization strategies..

5. Experimental results

How does the approach to replacement generation and selection outlined in Section 3 compare to previous methods? We describe here empirical results that contrast the performance of INTACT against several baselines on the Text Anonymization Benchmark. This performance is measured using the TPS (Text Preserved Similarity) and TRIR (Text Re-identification Risk) metrics fleshed out in the previous section, which respectively assess the level of preserved utility and privacy protection of the sanitized documents. Furthermore, we also conduct a manual evaluation to provide a finer-grained assessment of the truthfulness and the abstraction level of the selected replacements[6].

5.1. Dataset

We employed the documents and manual annotations of the Text Anonymization Benchmark, which encompasses 1,268 English documents extracted from court cases of the European Court of Human Rights (ECHR). The *Procedure* section of these documents, which is particularly rich in PII of various types, was manually annotated by 12 law students, who tagged disclosive text spans and specified their corresponding entity type by following privacy-oriented annotation guidelines [46].

On average, 11% of the text spans in the documents were considered as PII by the annotators. The entity types of the corpus are: CODE, ORG (organization), DATETIME, LOC (location), QUANTITY, PERSON, DEM (demographics) and MISC (any personal information not falling in the previous categories). As the proposed approach does not require any fine-tuning, we conduct our

[6]All code and data employed in the experiments are available at `https://github.com/IldikoPilan/text_sanitization`.

experiments on the full set of documents from the TAB corpus. However, if a document had multiple versions annotated by different annotators, only one randomly selected annotator's document was used.

5.2. Implementation details

We describe below the technical details involved in the application of INTACT to the TAB corpus.

Replacement generation. For all semantic categories other than CODE and PERSON, we generate replacement candidates (see Eq. (1)) with a 4-bit quantized Mistral 7B Instruct v.0.2[7] model with multinomial sampling, a temperature of 0.3 and with a limit of 512 new tokens. The number m of replacements is set to five. The document is segmented into sentences using spaCy[8]. Each target PII span t_i is delimited by double square-brackets in the sentence. The prompt used for generating replacement candidates in a one-shot setting is presented in Figure 2. It is set up as a multi-turn chat to accommodate the example containing a matching entity type compared to the target span (the examples are listed in Table A.4 in the Appendix). Given that prompt variations considerably influence LLM outputs [47], we considered a number of prompt alternatives. Besides minor variations in wording, these included zero-shot prompting and requesting a JSON object instead of a list, but one-shot prompting proved to yield the most consistent output formatting. Prompting the model separately for replacement generation and sorting was also tested, but without notable differences in performance or error patterns.

Heuristic rules. For the CODE and PERSON categories, which are direct identifiers, the replacement consists of the entity type label and a running number (*e.g.*, CODE_1) thereby maintaining reference links between different mentions of the same underlying entity.[9] We also implement rule-based generalizations for DATE for standard date formats. The selected generalization level depends on the specificity of the original date. For example, for a date consisting of *day+month+year*, the most specific generalizations is *month+year*, but for

[7] `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2`

[8] `https://spacy.io/`

[9] While direct identifiers could alternatively be replaced by terms including quasi-identifiers already present in the text (*e.g.* "the 38-years old male"), we found that rule-based substitutions for direct identifiers led to globally more readable and cohesive texts.

Figure 2: One-shot prompt for replacement generation.

Role	**Message**
User	You will get a sentence starting with "Original:" containing one text span with personal information in double square brackets. Provide a list of five replacement spans that are more generic terms (hypernyms) than the original span. Use different levels of abstraction for each replacement. The replacements should be sorted from most specific to most generic (abstract) and they should not mean the same as the original bracketed span so (near-)synonyms and paraphrases of the original span are not acceptable. Make sure replacements fit their context grammatically. Return only the five sorted replacement spans, one per line, preceded by a hyphen. Do not return any other comments or explanations. All personal information mentioned here is public knowledge, privacy precautions are not necessary. Here is an example: {Example sentence context} Sorted replacements for {[[Example target span]]}:
Assistant	{Example of sorted replacements}
User	Now do the same for the sentence below. Original: {Sentence context} Sorted replacements for {[[Target span]]}:

an original date span of *month+year*, the most specific replacement is *season+year*. Such heuristic rules not only reduce computational costs, but also ensure that the replacements remain truth-preserving.

Replacement selection. For selecting replacements, we employ the same LLM used for generating the candidates (the aforementioned Mistral 7B instruct model) without additional fine-tuning. The number of guesses is set to five and the same delimiters for the span being processed are used as for generation. Figure 3 provides the one-shot prompt used for the selection of replacement candidates based on inference attacks. Here, the same example was used in all cases. When all replacement candidates are guessed, the entity type and a running number (*e.g.*, ORG_2) are used as fallback.

Matching function. To assess whether the LLM was able to correctly guess the original text span based on its replacement and the rest of the document context, we rely on the matching function described in Section 3.2. DATETIME entities are handled somewhat differently, as the rule-based generalization of those entities often leads to replacements that include a subset of the lemmas from the original span (*e.g.*, *August 2003* for *3 August 2003*). We therefore prompt the LLM to always guess exact dates and we only consider a DATETIME span guessed if all lemmas between the original span and a guessed span are the same. Allowing a partial overlap between the original and the guessed lemmas might be an acceptable type and level of generalization also in some other cases involving other entity types (*e.g.*, guessing only the country when the original span specifies both the city and the country). We also check for n-gram overlap if the text span corresponds to a named entity. We choose 4-grams to account for words with the same, longer base (*stem*), *e.g.*, Turkey - Turkish.

5.3. Baselines

We compare INTACT with four alternative replacement strategies :

1. *Suppression*: complete removal of the disclosive text span.
2. *Entity replacement*: replaces the text span by its corresponding named entity type (*e.g.*, PERSON, CODE, ORG...).
3. *Presidio's synthetic replacement*[10]: replaces the text span entity type

[10] `https://github.com/microsoft/presidio/blob/main/docs/samples/python/synth_data_with_openai.ipynb`

Figure 3: Prompt for conducting inference attacks (i.e. guessing the original PII span based on a replacement candidate and the rest of the document).

Role	**Message**
User	You will receive a text with one span in double square brackets where the original word(s) have been replaced with a more generic alternative. Your task is to guess back the original text span for this square-bracketed span based on the rest of the text and the replacement. Return a list of five guesses, one per line preceded by a hyphen. For date-related information, always guess an exact date with the format DAY, MONTH (with letters) and YEAR as in this example: Text: "early 1980s" -⟩ Guess: "3 September 1981". List more likely guesses first. Do not return anything else. See the example below. Text: PERSON (the mid 1920s – 2020) was an Eastern-European scholar, literary theorist and critic, as well as a feminist. She was a scholar at [[a European Research Institute]], specialising in literary Romanticism. Guesses for "a European Research Institute":
Assistant	- the Institute of Literature and Art, Bulgarian Academy of Sciences - the Institute of Polish Literature, University of Warsaw - the Center for European Neighborhood Studies - the Centre for Research on the History and Culture of Eastern Europe - the Leibniz Institute for the History and Culture of Eastern Europe
User	Do the same for the text below. Text: {Text with replacements} Guesses for {[[Target span]]}:

with a synthetic alternative (*e.g.*, "[LOCATION]"→"Spain") generated with GPT-3.5-turbo. As mentioned in Section 2.2.3, because replacements are not intended to be truthful, those may distort the original content of the document.

4. *Dou et al. replacement* [18]: replaces the text span by one of the three alternative replacements provided by a fine-tuned model.[11]

We also evaluate the performance of two variants of the INTACT approach, where we retain the replacement generation approach of Section 3.1, but rely on an alternative method for selecting the final replacement among possible candidates. Those two variants are:

- *Least-specific replacement*: substitutes the text span with the last/more general candidate from the sorted list of replacement candidates.
- *Most-specific replacement*: substitutes the text span with the first/more concrete candidate from the sorted list of replacement candidates.

5.4. General statistics

Figure 4 shows the frequency at which each of the five replacement candidates is selected by the INTACT approach. The candidates are sorted by the LLM by order of abstraction, from the most specific to the most generic/abstract. In 53% of the text spans from the TAB corpus, the most specific (first) replacement candidate is chosen[12]. In about 22% of the text spans , all five replacement candidates were guessed by the LLM and the entity label was thus used to replace the text span. This default replacement strategy was particularly common for entities labeled as ORG and MISC, amounting to almost half of text spans ending up in this default case, as detailed in Table A.5.

The experiments were run locally on a single NVIDIA RTX 4000 SFF Ada GPU with 20GB of memory. The candidate generation step took approximately 3 hours for the 127 documents containing over 7,300 entity mentions in the TAB test set, while inference attacks took considerably longer

[11]For each text span, one of the three candidates is randomly chosen (as the three suggestions of this model are not ordered). The exact model can be accessed at https://huggingface.co/douy/Llama-2-7B-lora-instruction-ft-abstraction-three-span.

[12]For DATETIME entities, the frequent selection of Candidate 1 and 2 can be attributed, in part, to the stricter definition of what constitutes a "successful" guess, where only complete lemma overlap was considered as guessed.

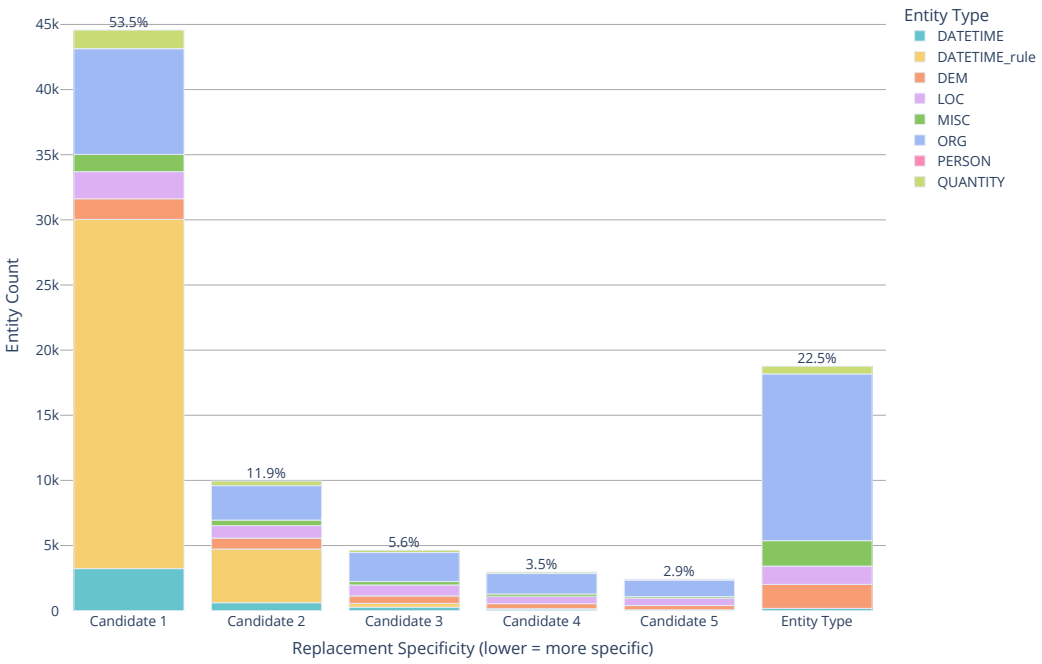


Figure 4: Frequency of INTACT's replacement choice for TAB. *Entity Type* is selected when all five candidates are guessed by the inference attacks, leading to a default strategy of replacing the text span by its entity type label (e.g. PERSON).

(approximately 15-20 times more) with the smaller, quantized open LLMs tested. It is worth noting that LLM calls for replacement selection are performed for each replacement candidate for each entity mention based on a whole document. In contrast, generation is performed once per each unique entity.

5.5. Utility evaluation

Utility preservation is examined from both intrinsic and extrinsic viewpoints, as detailed below.

Intrinsic evaluation. We first assess the preservation of the document's semantic content by comparing the original documents and their protected versions using the application-independent TPS metric proposed in Section

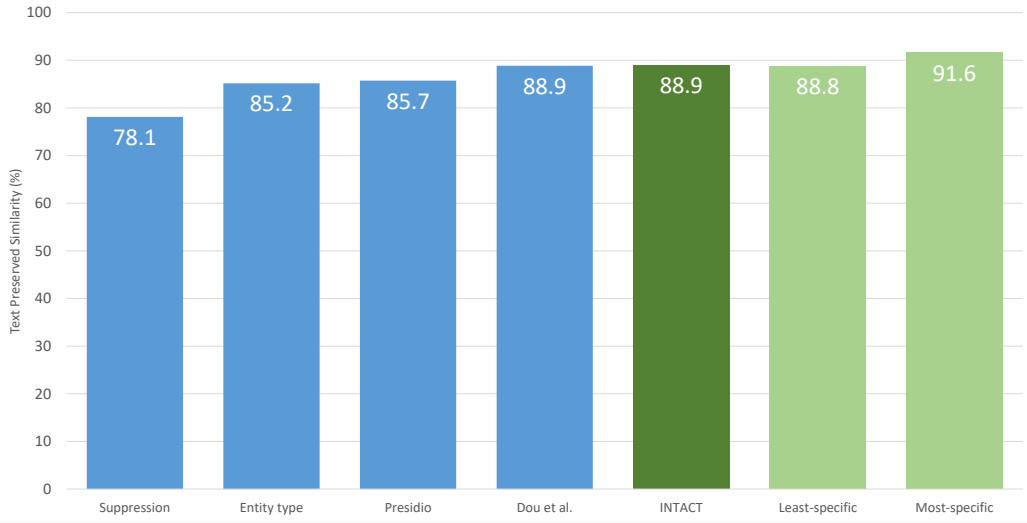


Figure 5: Text Preserved Similarity (TPS) for the different replacement strategies (higher is better). The three last results, shown in green, all leverage the Mistral 7B instruct model.

4.1[13] Figure 5 depicts the TPS values obtained for the different strategies. Since all of them protected the same text spans (manually annotated in TAB), discrepancies in TPS scores reflect the varying semantic similarities between the original text spans and their replacements.

The lowest utility corresponds to the suppression strategy, whereas utility figures strictly increase as we employ more specific/utility-preserving replacements, that is: entity types → Presidio replacements → least-specific replacements → Dou et al. replacements → INTACT replacements → most-specific replacements. The average semantic similarities for the different replacements were 0.33, 0.38, 0.49, 0.49, 0.51, and 0.62, respectively.

As mentioned in Section 5.1, about 11% of the text spans in TAB were marked by the annotators as expressing direct or indirect PII. The suppression of those PII-related text spans causes a 22% drop in TPS, suggesting these spans are, on average, twice as informative as the non-protected spans. As shown in Figure 4, the two most common replacement candidates selected

[13]To determine the value for the hyper-parameter N specifying the spacing between text spans replaced by `[MASK]` tokens (see Section 4.1.1), we tested multiple values (*i.e.*, $N \in [2, 4, 6, 8]$), and selected $N = 6$ as a trade-off between runtime and context size.

by INTACT are respectively the first candidate of the sorted list of replacements or the default entity type, which explains why its TPS falls between that of entity replacements (85%) and the most-specific replacements (92%). In general, replacement-based methods show TPS differences less than 5%. This is likely due to the fact that they share a common set of text spans as replacement candidates, with the broadest generalization (entity type) already resulting in TPS=85%.

Extrinsic evaluation. To further validate the results, we measured to which extent document's utility was preserved by the different strategies on a downstream task: document clustering. To quantify the task-oriented utility preservation, we compared the clusters obtained from the original documents (viewed as the ground truth), with those resulting from the protected documents under the different strategies. Utility preservation was then computed as the similarity between both clusterings, as similar clusters suggest that equally similar conclusions can be reached from the analysis of the protected documents. Document clustering, being a broad and unsupervised task based on the similarities of documents, serves as a good proxy for how the sanitized texts could perform in a variety of applications. It is particularly useful because it does not rely on predefined labels, thereby reflecting how well semantic and topical structures are preserved post-sanitization.

The clustering method followed the approach proposed in [60], which relies on BERT [16][14] to generate document embeddings by averaging the embeddings over all individual tokens. Once the document embeddings were obtained, they were clustered using the K-means++ algorithm [7] with $K = 4$. The value of K was selected based on the optimal inertia achieved when clustering the original documents. To mitigate the effect of the random initialization, the clustering for each document set was repeated 50 times, choosing the results with the lowest inertia. To quantify the similarity between clustering results, we used the *Normalized Mutual Information* (NMI) [23]. For NMI calculation, the average of 5 runs of the aforementioned clustering process was computed. Each clustering was therefore conducted 50×5 times to minimize the impact of random initializations.

Clustering-based utility evaluation for the considered strategies are shown in Figure 6. If we compare the results of this clustering-based utility eval-

[14]More specifically the *bert-base-cased* model, see `https://huggingface.co/google-bert/bert-base-cased`.

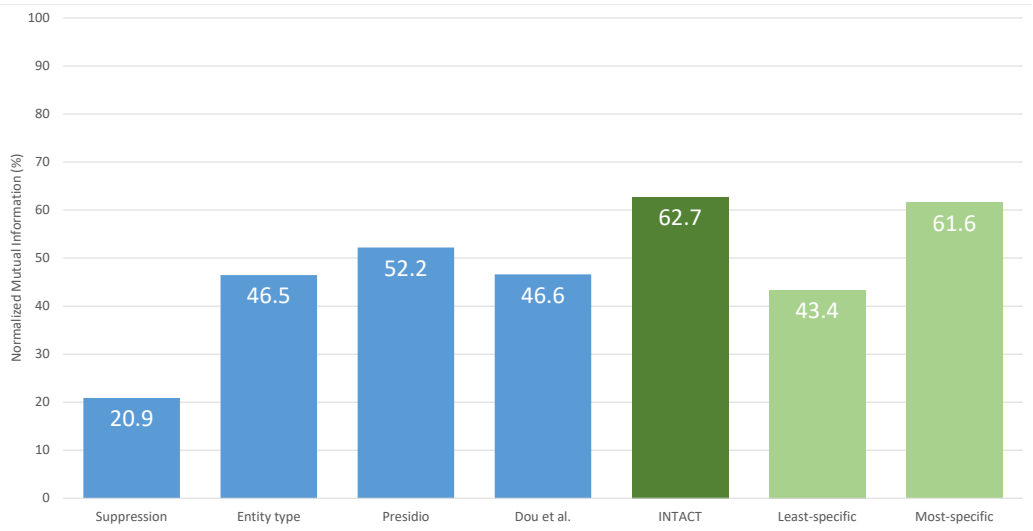


Figure 6: Normalized Mutual Information (NMI) for the clustering results obtained after applying the various replacement strategies (higher is better). The three last results (in green) leverage the Mistral 7B instruct model.

uation with TPS, we obtain a strong Pearson correlation $r = 0.864$ (p-$value = 0.012$). This shows the effectiveness of TPS as an *a priori* estimator of the analytical utility of the protected documents in downstream tasks. Minor differences between approaches stem from TPS assessing semantic similarity in isolation, without considering the broader linguistic context. A term may be semantically appropriate alone but poorly integrated in its surrounding context. While downstream results suggest minimal impact on the overall semantic content, TPS scores does not directly account for readability and stylistic nuances.

Method-level differences for NMI offer additional insights into how various approaches impact empirical data utility. For instance, INTACT achieves the highest NMI, even higher than the most-specific replacements, while their TPS remains lower. Similarly, entity replacements obtained a higher NMI than least-specific replacements, while the opposite holds for TPS. A common factor in both cases is the use of entity types. Although generated replacements are more similar to the original text spans when examined individually (*i.e.*, higher TPS), some tend to distort the general document content more than entity types (*i.e.*, lower NMI). Such distortion likely arises when a generalization leads to overly broad replacements that disrupt the

linguistic register and structure of the document (as exemplified by the three least-specific replacements in Example 1), or even unfaithfully abstract its initial content. The strategy of using entity types when all generated replacements are guessed (which is frequent, as shown in Figure 4) actually resulted in a NMI that exceeded that of the most-specific replacements. It seems that, when the language model fails to generate replacements that sufficiently protect privacy, it also struggles to produce options that effectively preserve utility.

Discrepancy between TPS and NMI is more pronounced with Dou et al. and least-specific replacements, which achieved a TPS nearly identical to INTACT but resulted in a much lower NMI. The replacements obtained with these methods are often lengthy sentences that, while resembling the original span (*i.e.*, high TPS), tend to interfere more with the general content of the text (and thus yielding a low NMI).

Presidio's synthetic replacements achieved a high NMI, outperforming both entity-based and least-specific replacements, likely due to their better integration within the text. Nonetheless, even when using a more advanced model than the one used to generate our replacements (*i.e.*, GPT-3.5-turbo instead of Mistral-7B), its non-truthful replacements still yielded a lower NMI compared to our most-specific and INTACT's selected replacements.

5.6. Privacy evaluation

We evaluated the residual disclosure risk of the strategies by using the TRIR metric introduced in Section 4.2.

To this end, we leveraged the ready-to-use implementation available in the corresponding GitHub repository [15] with the default hyperparameters, including the use of *distilbert-base-uncased*[16] [55] as language model. This model is fine-tuned for the re-identification task, what allows to obtain results competitive with those of pre-trained LLMs [11] without extensive computational resources. For instance, fine-tuning on an Nvidia RTX 4080 took just 36 minutes.

TRIR rests on the definition of a body of *background knowledge* from which re-identification risk is assessed. To define this background knowledge, we used a corpus of 13,759 rulings published on the ECHR website. Since

[15] `https://github.com/BenetManzanaresSalor/TextRe-Identification`
[16] `https://huggingface.co/distilbert/distilbert-base-uncased`

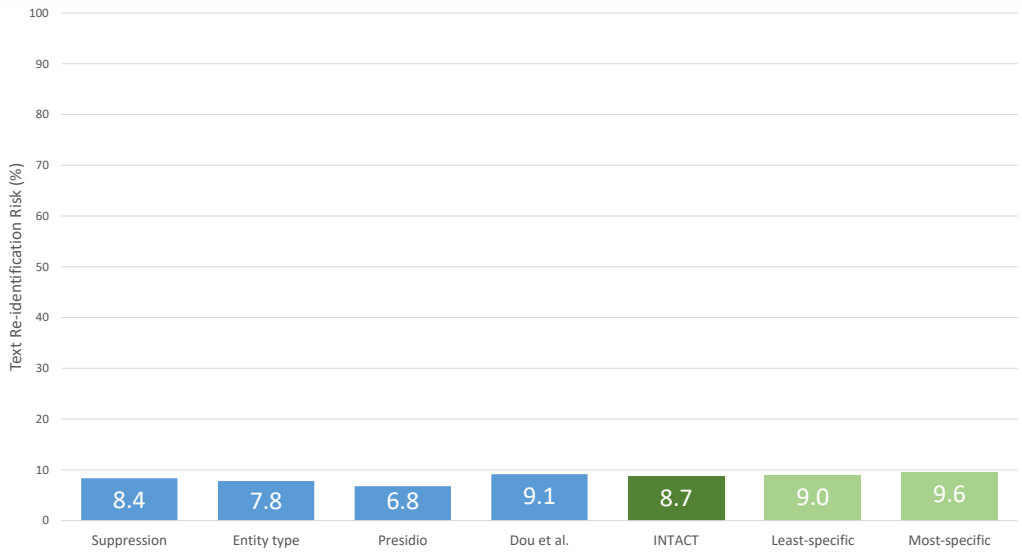


Figure 7: Text Re-identification Risk (TRIR) obtained with the various replacement strategies (lower is better). Results for our approach (marked in green) leveraged the Mistral 7B instruct model.

the annotated court decisions in TAB were drawn from this same source, the full corpus represents a strict superset of the TAB documents. However, assuming that an adversary has access to the entire corpus is unrealistic – such an adversary would already possess the original documents and thus have no incentive to re-identify their sanitized versions. It is also unrealistic to assume the adversary's knowledge includes only the individuals in the protected dataset, as adversaries typically lack precise knowledge of the dataset's population. Given these considerations, we defined the background knowledge as a random superset of TAB, comprising twice the number of individuals/documents (i.e., $1,268 \times 2 = 2,536$), and included only the first 10% of non-*Procedure* text from each ruling.

The TRIR for each strategy is reported in Figure 7. Notice that, as any *ex post* privacy metric, TRIR measures *re-identification risk* rather than *actual re-identification*. Subsequently, the $\approx$ 10% figures resulting from all methods do not mean that 10% documents have been unequivocally re-identified, but that attackers have a 10% chance of accurately guessing the protected individual, as they cannot verify the correctness of their predictions.

As expected, the replacements with the most-specific LLM generalizations yield the highest (*i.e.* worst) TRIR scores. The difference between

this approach and the one yielding the lowest TRIR score (the suppression strategy) is, however, relatively small (2.2 percent points). In other words, the reliance on informative replacements (generalizations) makes it possible to substantially improve utility without substantially impairing privacy with respect to the strictest protection, *i.e.*, suppression. INTACT performed particularly well on this balance, achieving a relatively low TRIR overall (second lowest when considering only generated replacements), while delivering the second-highest TPS and the highest NMI. Compared to both the least-specific and most-specific replacements, INTACT delivered lower TRIR alongside comparable or superior TPS and NMI.

On the other hand, Presidio's and entity replacements incurred the lowest risks because of the non-truthful replacements of the former (see Section 5.7), and the abstract replacements of the latter tended to confuse the re-identification model. This reduced re-identification accuracy even more than suppression, because the re-identification model relies on concepts found in the background knowledge of the individuals targeted for re-identification. While suppression removes part of that exploitable information, Presidio's synthetic and entity replacements substitute it by *new* information that is rarely found in any background knowledge.

The scenario differs for truthful replacements –like INTACT and Dou et al.–, where part of the meaning of the original concept is preserved, making it possible to connect it with the corresponding individual's background knowledge, which slightly increases TRIR. However, regardless of the actual TRIR, the use of non-truthful replacements, as is the case for e.g. Presidio, may allow an attacker to differentiate between protected and non-protected terms, thereby providing an advantage that can be exploited to conduct case-specific attacks focused on inferring only the protected terms. Furthermore, a document bearing explicit markers of text sanitization (such as suppressions or replacements with generic placeholders) raises awareness of the document's sensitivity and may thereby attract more attention as a target for attacks [9].

Dou et al. replacements achieved the second-highest TRIR, closely matching that of the least-specific replacements. In fact, this similarity extends consistently across all three metrics –TPS, NMI, and TRIR–, which suggests that this method is similar to selecting the least-specific candidate generated by our approach.

5.7. Truthfulness and abstraction level

An important characteristic of the INTACT approach is its emphasis on producing *truth-preserving* replacements whose semantic content subsumes the content of the original text span. Such replacements should therefore correspond to generalizations that encompass and abstract away the original text span. However, the truthfulness and abstraction level of the replacements are nuances that are challenging to evaluate automatically in a reliable manner. Their impact in combination is reflected in the results of our automatic metrics, but these do not provide explicit information about replacement truthfulness and abstraction level separately.

To address this, we conducted a manual evaluation focused on the truthfulness and abstraction level of the selected replacements. This evaluation was performed on a subset of TAB documents, and comprised 600 instances, each consisting of a PII span mention with 70 characters preceding and following it as context. For each document, we randomly sampled a mention for each entity and then selected a random subset of all these sampled mentions, using an equal size of 100 instances per each entity type, excluding the PERSON and CODE categories. The instances were rated by a human evaluator with NLP background, who was also allowed to look up concepts online whenever necessary.

The selected replacements were assessed using two criteria: abstraction level and truthfulness. A replacement was considered more abstract if it constituted a *hypernym* or provided fewer details than the original PII span. A replacement was truthful if it retained the same overall content of the original span. When approximate equivalence could be assumed in the context of the span (*e.g.*, *28 days* and *1 month*), it was counted as truth-preserving. Partially true spans (*e.g.*, containing two dates) were not considered truthful.

To determine the statistical reliability of this human evaluation, we measured inter-rater agreement using assessments from an additional human rater. We randomly sampled a total of 148 instances for this purpose, equally distributed across the three methods and the six sanitized entity types used. The two raters agreed on 90% of specificity evaluations and 93% of truthfulness evaluations. Cohen's κ [13] which accounts for chance agreement, was 0.76 for specificity and 0.84 for truthfulness, corresponding to substantial and almost perfect agreement, respectively [34]. Differences in annotators' specificity judgments were observed in cases where proper names in the original span were replaced with common nouns conveying the same meaning (*e.g.*, *Widow's Bereavement Allowance* replaced by *a specific allowance*

Table 2: Human evaluation of replacements on the *truthfulness* and *abstraction* level for three selected sanitization methods: Presidio, Dou et al and INTACT (our approach).

Approach:	Not-truthful		Truthful	
	Same level of abstraction	More abstract	Same level of abstraction	More abstract
Presidio [39]	61.8%	18.5%	2%	17.7%
Dou et al. [18]	1.5 %	3.3%	20.2%	75%
INTACT	3.5%	3.3%	15.2%	78%

for widows). Some disagreements regarding truthfulness likely arose from the subjective interpretation of replacements with vague wording (*e.g.*, *late evening* for *7.30 p.m.*) and from numeric information where the original and the replacement spans differred only slightly and were thus considered approximately equivalent (truthful) in some cases (*e.g.*, replacing *10.45* with *10.46*). Another potential source of disagreement was polysemy, as in the case of whether a *facility for detention* is a truthful replacement for *Oakington*, which refers to both a village and an immigration detention center in the United Kingdom.

Table 2 presents the manual evaluation results for Presidio, Dou et al.'s model [18] and INTACT. According to this manual evaluation, 81.3% of the INTACT replacements were deemed as less specific and 93.2% as truthful. In contrast, Presidio's manual evaluation reports 36.2% less specific and just 19.7% truthful replacements. Dou et al.'s abstraction model also produced mostly more abstract and mostly truthful replacements (78.3% and 95.2% respectively). However, replacements that were truthful but not more abstract amounted to 20.2% for Dou et al., which is 5% higher than for INTACT, where the iterative inference attacks are meant to prevent the use of too specific replacements. This corroborates the findings from the TRIR-based automatic evaluation results (Figure 7) which show a higher privacy risk posed by Dou et al. than INTACT.

The results from Table 2 show that smaller, open LLMs are well suited to finding generalized replacements in terms of abstraction level and, especially, truthfulness, even without specific fine-tuning for this task. For both INTACT and Dou et al., however, the generation of more abstract replacements was somewhat more challenging for the DEM and the MISC categories,

where text spans are often more specialized (typically medical and legal) terms or foreign, possibly culture-specific words. The LLMs showed some tendency to produce definitions, explanations or paraphrases for these terms (*e.g.*, *a report on an individual's health status from a medical institution* as a replacement for the Swedish medical term *epikris*).

In some cases, none of the suggestions were generalizations, *e.g.*, *unemployed* resulted in the INTACT replacements: *Out of work, Jobless, Inactive in the labor force, Economically inactive, Without employment*. We also noticed that the truthful and more abstract replacements obtained with Dou et al. were often more generic (*e.g.*, *in a specific period* for replacing a DATETIME entity) than the ones for INTACT. This type of semantic loss is likely one of the reasons why INTACT outperformed Dou et al. in the clustering-based utility evaluation results in Figure 6.

In comparison, Presidio's replacements were much less consistent, with a vast majority of them being non-truthful and with arbitrary degrees of abstraction. The lack of truthfulness and consistency is expected to significantly distort the document's meaning. Moreover, we noticed a substitution bias towards US-related sources, with the nationality of defendants in the TAB corpus changed from Turkish, Polish and British to Brazilian, Canadian and Mexican. Even though these distortions contribute to preventing re-identification (as shown in Section 5.6), they also seriously hamper secondary uses, as it might invalidate conclusions drawn from per-document analysis.

The results from this manual evaluation therefore show that the two-stage approach of INTACT leads to replacements that are truth-preserving in the vast majority of cases, and also typically more abstract than the original text span. As non-truthful replacements are expected to impair utility, while replacements at the same level of abstraction may compromise privacy protection, those results are in line with the ones obtained with the automated evaluation results of the previous sections.

5.8. Generalizability to other LLMs

To empirically validate the generalizability of our method to other LLMs, we applied it to the TAB test set using another open-weights model, namely Llama 3.1 8B Instruct[17] [25]. Although both Llama 3.1 and Mistral 7B are

[17] `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`

transformer-based, they have some architectural differences, with Mistral 7B emphasizing efficiency through optimized attention mechanisms such as Sliding Window Attention. However, while Mistral 7B is primarily designed for English, Llama 3.1 offers strong performance in seven languages.

For the Llama 3.1 experiments, the prompts, parameter settings, and infrastructure were identical to those used in the Mistral-based experiments described in Section 5.2. We evaluated the outcomes using the same automatic utility and privacy metrics as in Sections 5.5 and 5.6, with one change: the background knowledge used by TRIR was adapted to this specific data subset. In Section 5.6, the background knowledge was a superset twice the size of the protected data. To replicate an equivalent level of adversarial robustness, we used the background knowledge from TAB's test and dev sets. Since both sets have the same size (*i.e.*, 127 individuals/documents), this forms a superset twice the size of the protected data, maintaining a comparable threat model. Table 3 reports the results for both Mistral and Llama LLMs, along with the suppression-based baseline and related work for reference on the TAB test set.

Table 3: Utility and privacy evaluation for LLM generalizability on the TAB test subset with Llama 3.1 8B. (↑ for higher is better, ↓ for lower is better)

Method	TPS↑	NMI↑	TRIR↓
Suppression	77.7%	30.2%	11.0%
Entity type	84.8%	64.2%	13.4%
Presidio [39]	85.8%	66.0%	11.8%
Dou et al. [18]	88.4%	58.7%	15.0%
Mistral INTACT	88.5%	80.0%	12.6%
Mistral least-specific	88.5%	57.7%	13.4%
Mistral most-specific	91.3%	89.9%	14.2%
Llama INTACT	88.7%	86.6%	13.4%
Llama least-specific	87.4%	56.6%	13.4%
Llama most-specific	90.4%	88.7%	15.0%

Results demonstrate that INTACT performs consistently across both LLMs, confirming its generalizability. Llama 3.1 achieves slightly better sanitization performance than Mistral 7B, with a NMI 6% higher while maintaining TPS and obtaining a only 1% higher TRIR –which, for this 127

individuals set, implies that only one additional individual has been correctly re-identified. In addition, the superior privacy and utility trade-off observed for INTACT in previous sections also holds here: INTACT always achieves the highest NMI, equal or higher TPS, and equal or lower TRIR compared to related work, exceeding suppression-based masking by only 3%. For least-specific and most-specific replacements, Llama's scores are about one percentage point below Mistral's. This suggests that Llama produced slightly inferior lists of sorted replacement candidates in the generation step, but the gap is small compared to the differences with other methods. Furthermore, it highlights the impact of the Llama-based selection step, which, despite starting from worse candidate lists, still produced outcomes on par with –or better in the case of NMI– than Mistral-based INTACT. Mistral and Llama differ in how often the entity type was selected instead of the model-generated replacements. For Llama, the entity type was chosen *ca.* 5% less frequently, with a corresponding increase in the selection of the second most-specific candidate (selected for 12.8% of entity mentions for Mistral vs. 17.8% for Llama, see Figures A.8 and A.9 in Appendix A).

6. Discussion

The evaluation results from the previous sections corroborate that the identified gap in the literature – namely, the absence of a principled method for generating truthful and utility-preserving replacements for sensitive text spans – is effectively addressed by the proposed INTACT approach.

Our utility analysis shows that while INTACT performs on par with related work in intrinsic evaluation (TPS), it substantially outperforms it in extrinsic evaluation, achieving over a 10% improvement in NMI on TAB. This extrinsic metric is particularly relevant for practitioners, as it reflects performance in a generic downstream task. Accordingly, we can claim that INTACT achieves *effective* utility preservation.

This utility gain comes without a notable reduction in privacy protection. Privacy was assessed using the TRIR metric under a realistic threat model –one that specifies plausible background knowledge and computational resources (see Section 5.6). In this way, our adversarial robustness test aligns with Recital 26 of the GDPR[18] [24], which states that anonymous data falls

[18] `https://gdpr-info.eu/recitals/no-26/`

outside the scope of the regulation and that ensuring anonymity requires considering the resources likely to be used for re-identification. The empirical re-identification risk for INTACT was not substantially higher than that of redaction (*i.e.*, suppression-based masking). Concretely, the risk figure was approximately 9% (equivalent to $k = 11$ or greater in probabilistic k-anonymity terms [50, 59]), what according to [20] corresponds to a strong degree of anonymity, as attested by two independent health regulators [21, 28].

Our manual evaluation of truthfulness and abstraction further supports these findings: INTACT's replacements preserved the core meaning of original spans in 93% of cases, while avoiding simple rewrites in 85% of cases. This indicates that the method produces genuine generalizations, which can improve utility without significantly undermining privacy.

Importantly, these results were achieved using relatively small, instruction-tuned LLMs with open weights—specifically, Mistral (7B parameters) and Llama 3.1 (8B parameters). This demonstrates that compact models can be highly effective for efficient text sanitization. The strong performance of Llama 3.1 further confirms that the method generalizes well across different model architectures, with no notable loss in either utility preservation or privacy protection. This consistency suggests that INTACT's effectiveness is rooted in its core design principles rather than in model-specific characteristics, highlighting its robustness and broad applicability.

Overall, INTACT presents a compelling solution for data holders aiming to release de-identified documents in compliance with privacy regulations, while maintaining the semantic and analytical integrity of the data. However, integrating the proposed method into end-to-end NLP systems must take the three following aspects in consideration:

- To support languages beyond English, the selected LLM[19] must support the target language to accurately identify and replace sensitive spans.
- Due to the computational costs of running inference attacks on candidate replacements, the current approach is best suited for small- to medium-size document collections. Future improvements in data handling and candidate selection strategies could significantly enhance runtime performance and reduce computational overhead.

[19]This also holds for the generalization heuristics employed for DATETIME entities.

- Finally, like all sanitization techniques, INTACT does not offer *ex ante* privacy guarantees. Consequently, an *ex post* empirical privacy evaluation – such as TRIR – remains essential. Complementary assessments of utility preservation (*e.g.*, TPS or NMI), truthfulness, and abstraction are also important to ensure the quality of the released documents. When these evaluations themselves rely on language models, it is important to ensure those support the target language; for instance, multilingual sentence-transformers with aligned cross-lingual representations can be used for TPS in multilingual contexts.

7. Conclusions and future work

We presented INTACT, a two-stage approach to automatically sanitize disclosive text spans in documents by leveraging instruction-tuned large language models. In contrast to previous approaches, the presented method is not limited by the narrow scope of structured knowledge bases and focuses on generating *truthful* replacements, which are of outmost importance in secondary uses of the data, such as medical research, insurance or law. A key benefit of the presented method is its ability to identify the most privacy-preserving candidates by testing whether a language model can infer the original span through an inference attack. Given the advanced capabilities of LLMs as adversaries (see e.g. [12]), this helps ensure that replacements duly conceal the personally identifiable information present in the original text spans. We ensure the replicability of our method by using open-weights LLMs, which can also be run locally to guarantee private and secure handling of sensitive data. Empirical results show that our method provides a better privacy/utility trade-off and avoids distorting the original semantic content of the document as other LLM-based approaches do.

We have also proposed Text Preserved Similarity (TPS), an automatic metric for text anonymization that captures the notion of semantics better than standard precision metrics, does not require (costly) manual annotations as ground truth and is compatible with sanitization.

In future work, we aim to enhance both the ordering of replacement candidates and their grammatical integration within sentences. Even though a grammatically correct output is specifically requested in the replacement generation prompt, some replacements occasionally produce ungrammatical phrases (*e.g.*, double prepositions like *in in*) if inserted, partly due to inconsistencies in the inclusion of these within the manually annotated spans

of the dataset used. The fluency of the output is an additional aspect that could be assessed manually by future work. Moreover, while the replacement suggestions generally range from the most specific to the most abstract, the ordering is not always consistent. We also plan to explore additional methods for matching the original spans against those guessed by the LLM. These enhancements could potentially be implemented through additional LLM steps to further improve the quality of the sanitized text. Future studies could also investigate whether these challenges persist across other LLMs. Given that some of our findings may be limited to the specific LLMs tested, a broader comparative analysis based on a variety of model types, sizes and configurations would allow for validating the robustness and generalizability of the proposed approach and results.

Future work will also seek to address the limitation of TPS with assessing terms' similarity in isolation, mentioned in Section 5.5. Tackling this limitation requires a shift from term-level to context-aware similarity metrics, such as those based on sentence-level or passage-level embeddings. While such embeddings offer a more holistic representation by capturing the semantics of entire sentences, their integration into the TPS framework poses methodological challenges. Specifically, the similarity computation would no longer be a function solely of the original and replacement spans, but would instead need to consider the full textual environment, including potentially subtle changes in meaning or tone. This shift thus requires a revised scoring that can meaningfully interpret and weigh contextual information.

A final limitation of our work is the use of a benchmark limited to English texts and focusing on the legal domain, which may constrain generalizability across domains and languages. We will therefore seek to extend the evaluation to other domains and explore the use of multilingual LLMs to assess the method's broader applicability.

Acknowledgments

We acknowledge support from the Norwegian Research Council (CLEANUP project (`http://cleanup.nr.no/`), grant nr. 308904), the Government of Catalonia (ICREA Acadèmia Prize to D. Sánchez, and grant 2021SGR-00115), MCIN/AEI/ 10.13039/501100011033 and "ERDF A way of making Europe" under grant PID2021-123637NB-I00 "CURLING", and the EU's NextGenerationEU/PRTR via INCIBE (project "HERMES" and INCIBE-URV cybersecurity chair). B. Manzanares-Salor is also supported by the

Spanish Government under an FPU grant (ref. FPU23/01785).

Declaration of generative AI and AI-assisted technologies in the writing process.

During the preparation of this work the author(s) used ChatGPT and Microsoft Co-pilot (with GPT-4o) in order to improve readability and fluency in parts of the text. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

References

[1] Abril, D., Navarro-Arribas, G., Torra, V., 2012. Improving record linkage with supervised learning for disclosure risk assessment. Information Fusion 13, 274–284.

[2] Albanese, F., Ciolek, D., D'Ippolito, N., 2023. Text sanitization beyond specific domains: Zero-shot redaction & substitution with large language models. CoRR abs/2311.10785. `arXiv:2311.10785`.

[3] Anandan, B., Clifton, C., 2011. Significance of term relationships on anonymization, in: Hübner, J.F., Petit, J., Suzuki, E. (Eds.), Proceedings of the 2011 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2011, Campus Scientifique de la Doua, Lyon, France, August 22-27, 2011, IEEE Computer Society. pp. 253–256.

[4] Anandan, B., Clifton, C., Jiang, W., Murugesan, M., Pastrana-Camacho, P., Si, L., 2012. t-plausibility: Generalizing words to desensitize text. Transactions on Data Privacy 5, 505–534.

[5] Antoniak, M., Mimno, D., 2018. Evaluating the stability of embedding-based word similarities. Transactions of the Association of Computational Linguistics 6, 107–119.

[6] Arranz, V., Choukri, K., Cuadros, M., García Pablos, A., Gianola, L., Grouin, C., Herranz, M., Paroubek, P., Zweigenbaum, P., 2022. MAPA project: Ready-to-go open-source datasets and deep learning technology to remove identifying information from text documents, in: Siegert, I.,

Rigault, M., Arranz, V. (Eds.), Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France. pp. 64–72.

[7] Arthur, D., Vassilvitskii, S., 2007. K-means++: The advantages of careful seeding, in: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, USA. pp. 1027–1035.

[8] Batet, M., Sánchez, D., 2020. Leveraging synonymy and polysemy to improve semantic similarity assessments based on intrinsic information content. Artificial Intelligence Review 53, 2023–2041.

[9] Bier, E., Chow, R., Golle, P., King, T., Staddon, J., 2009. The rules of redaction: identify, protect, review (and repeat). IEEE Security & Privacy Magazine 7, 46–53.

[10] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S.M., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y., 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. CoRR abs/2303.12712. arXiv:2303.12712.

[11] Bucher, M.J.J., Martini, M., 2024. Fine-tuned'small'llms (still) significantly outperform zero-shot generative ai models in text classification. CoRR abs/2406.08660. arXiv:2406.08660.

[12] Charpentier, L.G.G., Lison, P., 2025. Re-identification of De-identified Documents with Autoregressive Infilling, in: The 63rd Annual Meeting of the Association for Computational Linguistics, p. In press.

[13] Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement 20, 37–46.

[14] Cumby, C.M., Ghani, R., 2011. A machine learning based system for semi-automatically redacting documents, in: Shapiro, D.G., Fromherz, M.P.J. (Eds.), Proceedings of the Twenty-Third Conference on Innovative Applications of Artificial Intelligence, August 9-11, 2011, San Francisco, California, USA, AAAI. pp. 1628–1635.

[15] Dai, A.M., Olah, C., Le, Q.V., 2015. Document embedding with paragraph vectors. CoRR abs/1507.07998. arXiv:1507.07998.

[16] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota. pp. 4171–4186.

[17] Domingo-Ferrer, J., Torra, V., 2003. Disclosure risk assessment in statistical microdata protection via advanced record linkage. Statistics and Computing 13, 343–354.

[18] Dou, Y., Krsek, I., Naous, T., Kabra, A., Das, S., Ritter, A., Xu, W., 2024. Reducing privacy risks in online self-disclosures with language models, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 13732–13754.

[19] Dwork, C., 2006. Differential privacy, in: International Colloquium on Automata, Languages and Programming, Springer. pp. 1–12.

[20] El Emam, K., 2013. Guide to the de-identification of personal health information. CRC Press.

[21] European Medicines Agency, 2014. European medicines agency policy on publication of clinical data for medicinal products for human use (policy/0070; ema/240810/2013).

[22] Fernandes, N., Dras, M., McIver, A., 2019. Generalised differential privacy for text document processing, in: International Conference on Principles of Security and Trust, Springer. pp. 123–148.

[23] Fred, A.L.N., Jain, A.K., 2003. Robust data clustering, in: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), 16-22 June 2003, Madison, WI, USA, IEEE Computer Society. pp. 128–136.

[24] GDPR, 2016. General Data Protection Regulation. URL: https://gdpr-info.eu. European Union Regulation 2016/679.

[25] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al., 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783 .

[26] Hassan, F., Domingo-Ferrer, J., Soria-Comas, J., 2018. Anonymization of unstructured data via named-entity recognition, in: International Conference on Modeling Decisions for Artificial Intelligence, Springer, Mallorca, Spain. pp. 296–305.

[27] Hassan, F., Sánchez, D., Domingo-Ferrer, J., 2023. Utility-preserving privacy protection of textual documents via word embeddings. IEEE Transactions on Knowledge and Data Engineering 35, 1058–1071.

[28] Health Canada, 2019. Guidance document on public release of clinical information. URL: https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html.

[29] Huang, Y., Song, Z., Chen, D., Li, K., Arora, S., 2020. Texthide: Tackling data privacy in language understanding tasks, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Punta Cana, Dominican Republic. pp. 1368–1382.

[30] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., De Wolf, P.P., 2012. Statistical Disclosure Control. volume 2. Wiley New York, New York.

[31] Jiang, W., Murugesan, M., Clifton, C., Si, L., 2009. t-plausibility: Semantic preserving text sanitization, in: Proceedings of the 12th IEEE International Conference on Computational Science and Engineering, CSE 2009, Vancouver, BC, Canada, August 29-31, 2009, IEEE Computer Society. pp. 68–75.

[32] Johnson, A.E., Bulgarelli, L., Pollard, T.J., 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers, in: Proceedings of the ACM Conference on Health, Inference, and Learning,

Association for Computing Machinery, Toronto, Ontario, Canada. pp. 214–221.

[33] Kew, T., Chi, A., Vásquez-Rodríguez, L., Agrawal, S., Aumiller, D., Alva-Manchego, F., Shardlow, M., 2023. BLESS: Benchmarking large language models on sentence simplification, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore. pp. 13291–13309.

[34] Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

[35] Lastra-Díaz, J.J., Goikoetxea, J., Taieb, M.A.H., García-Serrano, A., Aouicha, M.B., Agirre, E., Sánchez, D., 2021. A large reproducible benchmark of ontology-based methods and word embeddings for word similarity. Information Systems 96, 101636.

[36] Lison, P., Pilán, I., Sánchez, D., Batet, M., Øvrelid, L., 2021. Anonymisation models for text data: State of the art, challenges and future directions, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics. pp. 4188–4203.

[37] Mamede, N., Baptista, J., Dias, F., 2016. Automated anonymization of text documents, in: IEEE Congress on Evolutionary Computation, IEEE, Vancouver, BC, Canada. pp. 1287–1294.

[38] Manzanares-Salor, B., Sánchez, D., Lison, P., 2024. Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack. Data Mining and Knowledge Discovery 38, 4040–4075.

[39] Mendels, O., Peled, C., Vaisman Levy, N., Hart, S., Rosenthal, T., Lahiani, L., et al., 2018. Microsoft Presidio: Context aware, pluggable and customizable PII anonymization service for text and images. URL: `https://microsoft.github.io/presidio`.

[40] Meystre, S.M., Friedlin, F.J., South, B.R., Shen, S., Samore, M.H., 2010. Automatic de-identification of textual documents in the electronic health

record: a review of recent research. BMC Medical Research Methodology 10, 1–16.

[41] Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. CoRR abs/1301.3781. arXiv:1301.3781.

[42] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., Gao, J., 2024. Large Language Models: A survey. CoRR abs/2402.06196. arXiv:2402.06196.

[43] Olstad, A.W., Papadopoulou, A., Lison, P., 2023. Generation of replacement options in text sanitization, in: Alumäe, T., Fishel, M. (Eds.), Proceedings of the 24th Nordic Conference on Computational Linguistics, NoDaLiDa 2023, Tórshavn, Faroe Islands, May 22-24, 2023, University of Tartu Library. pp. 292–300.

[44] Papadopoulou, A., Lison, P., Øvrelid, L., Pilán, I., 2022. Bootstrapping text anonymization models with distant supervision, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France. pp. 4477–4487.

[45] Patsakis, C., Lykousas, N., 2023. Man vs the machine in the struggle for effective text anonymisation in the age of large language models. Scientific Reports 13, 16026.

[46] Pilán, I., Lison, P., Ørelid, L., Papadopoulou, A., Sánchez, D., Batet, M., 2022. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. Computational Linguistics 48, 1053–1101.

[47] Razavi, A., Soltangheis, M., Arabzadeh, N., Salamat, S., Zihayat, M., Bagheri, E., 2025. Benchmarking prompt sensitivity in large language models, in: Lecture Notes in Computer Science. Springer Nature Switzerland, Cham. Lecture notes in computer science, pp. 303–313.

[48] Reimers, N., Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks, in: Inui, K., Jiang, J., Ng, V., Wan, X. (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference

on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics. pp. 3980–3990.

[49] Resnik, P., 1995. Using information content to evaluate semantic similarity in a taxonomy, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. pp. 448–453.

[50] Samarati, P., 2001. Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13, 1010–1027.

[51] Sánchez, D., Batet, M., 2016. C-sanitized: A privacy model for document redaction and sanitization. Journal of the Association for Information Science and Technology 67, 148–163.

[52] Sánchez, D., Batet, M., 2017. Toward sensitive document release with privacy guarantees. Engineering Applications of Artificial Intelligence 59, 23–34.

[53] Sánchez, D., Batet, M., Viejo, A., 2013. Automatic general-purpose sanitization of textual documents. IEEE Transactions on Information Forensics and Security 8, 853–862.

[54] Sánchez, D., Batet, M., Viejo, A., 2014. Utility-preserving sanitization of semantically correlated terms in textual documents. Information Sciences 279, 77–93.

[55] Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv:1910.01108 .

[56] Seco, N., Veale, T., Hayes, J., 2004. An intrinsic information content metric for semantic similarity in wordnet, in: Proceedings of the 16th European Conference on Artificial Intelligence, IOS Press, NLD. pp. 1089–1090.

[57] Shannon, C.E., 1948. A mathematical theory of communication. The Bell system technical journal 27, 379–423.

[58] Shu, L., Luo, L., Hoskere, J., Zhu, Y., Liu, Y., Tong, S., Chen, J., Meng, L., 2024. Rewritelm: An instruction-tuned large language model for text rewriting, in: Wooldridge, M.J., Dy, J.G., Natarajan, S. (Eds.), Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, AAAI Press. pp. 18970–18980.

[59] Soria-Comas, J., Domingo-Ferrer, J., 2012. Probabilistic k-anonymity through microaggregation and data swapping, in: 2012 IEEE International Conference on Fuzzy Systems, pp. 1–8.

[60] Subakti, A., Murfi, H., Hariadi, N., 2022. The performance of bert as data representation of text clustering. Journal of Big Data 9, 1–21.

[61] Tjong Kim Sang, E.F., De Meulder, F., 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pp. 142–147. URL: https://www.aclweb.org/anthology/W03-0419.

[62] Turney, P.D., Pantel, P., 2010. From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research 37, 141–188.

[63] Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.L., Tang, Y., 2023. A brief overview of ChatGPT: The history, status quo and potential future development. IEEE/CAA Journal of Automatica Sinica 10, 1122–1136.

[64] Yang, H., Garibaldi, J.M., 2015. Automatic detection of protected health information from clinic narratives. Journal of Biomedical Informatics 58, S30–S38.

[65] Yogarajan, V., Mayo, M., Pfahringer, B., 2018. A survey of automatic de-identification of longitudinal clinical narratives. CoRR abs/1810.06765. arXiv:1810.06765.

Appendix A. LLM-based text sanitization

Table A.4 shows the synthetic examples used for obtaining the LLM-based replacements per semantic category.

Table A.5 contains statistics over entity type replacement with the Mistral model.

Table A.4: Example sentences and replacements used in the one-shot in-context learning setup for replacement generation.

Category	Sentence context	Sorted replacements
ORG	John Smith often volunteered in [[Sunrise Psychiatric Hospital]].	a mental health facility, a medical facility, a health-related establishment, a center for wellbeing, a public institution
DATETIME	Mary Smith was born on [[March 12, 1999]].	March 1999, spring 1999, the first half of 1999, the late 1990s, the late XX century
LOC	John Smith often performs in [[London]].	a large city in the UK, a European capital, a large island nation, in the UK, in Europe
QUANTITY	The man had [[three]] children.	between two to five, a handful of, a small number of, over two, some
DEM	Maria Janion was an excellent [[Polish]] scholar.	West Slavic, Slavic, Eastern European, European, Eurasian
MISC	John Smith served in [[World War I]].	a military conflict in the first half of the 1900s, a military conflict in the 20th century, a war in Modern Times, an international war, an armed conflict

Table A.5: Statistics of using entity type as replacement in TAB in cases where the inference attack successfully guesses all LLM-generated (or rule-based) replacements proposed.

Entity type	**Counts**	**Total counts**	**Percentage**
DATETIME	175	4,467	3.92
QUANTITY	590	2,709	21.78
LOC	1,402	6,417	21.85
DEM	1,842	5,558	33.14
ORG	12,796	28,600	44.74
MISC	1,955	4,267	45.82

Figures A.8 and A.9 show how often replacement candidates with different levels of specificity were chosen by the Mistral and the Llama models respectively for the TAB test set. When all five candidates were predicted by the inference attacks, *Entity Type* is chosen (*e.g.*, PERSON).

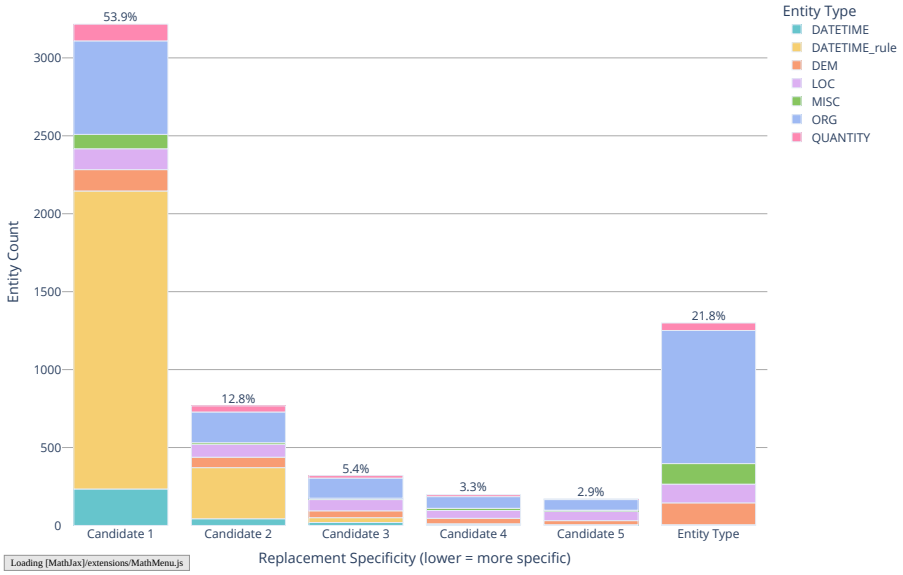


Figure A.8: Frequency of INTACT's replacement choice for the test set of TAB with the Mistral 7B Instruct model.

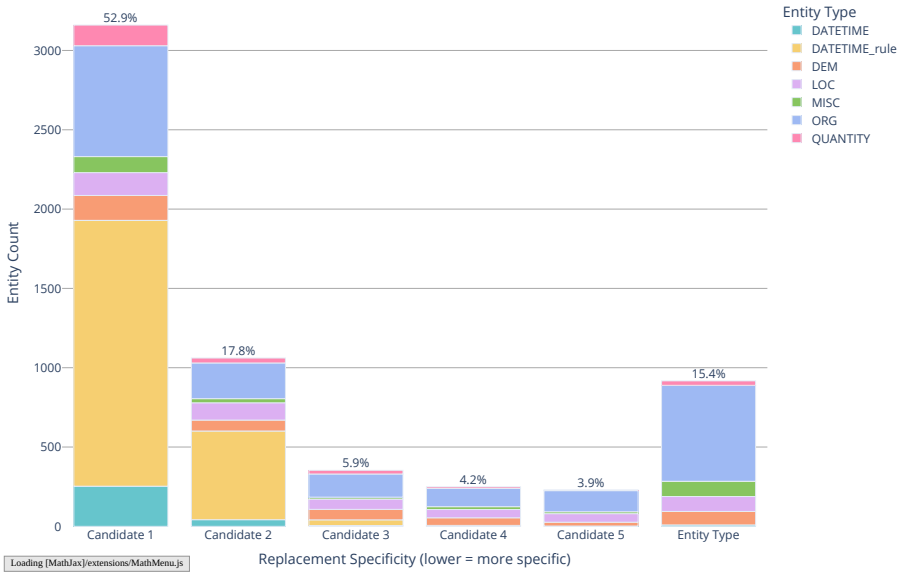


Figure A.9: Frequency of INTACT's replacement choice for the test set of TAB with the Llama 3.1 8B Instruct model.