A comparative analysis, enhancement and evaluation of text anonymization with pre-trained Large Language Models

Benet Manzanares-Salor[1] and David Sánchez[1]

Department of Computer Engineering and Mathematics, CYBERCAT-Center for Cybersecurity Research of Catalonia, Universitat Rovira i Virgili, Tarragona, 43007, Spain {benet.manzanares,david.sanchez}@urv.cat

Abstract. Large Language Models (LLMs) have gained prominence for their remarkable proficiency across various natural language processing tasks. Recent studies have suggested their potential to outperform current text anonymization methods, although an objective evaluation is needed to validate these claims. To address this issue, this work introduces a comprehensive evaluation framework that automatically assesses both privacy protection and utility preservation without relying on manually curated ground truth data. Moreover, we conduct an in-depth analysis of the LLM-based text anonymization methods proposed so far. Building on the strengths and limitations we found, we propose a novel method to enhance the anonymization quality. We also report extensive experimental comparisons between LLM-based approaches and a variety of previous techniques, including those based on named entity recognition (NER), and those more oriented towards privacy-preserving data publishing (PPDP). Results show that LLM-based approaches effectively outperform traditional methods both in terms of privacy and utility. Furthermore, we benchmark against manual anonymization, which performed poorly, thus highlighting the limitations of using them as evaluation ground truth. Notably, our LLM-based method stood out by achieving the best privacy protection, and the best privacy-utility trade-off.

Keywords: Text anonymization · Large Language Models · Privacy · Utility · Evaluation

1 Introduction

In an era where data serves as the cornerstone of innovation and decision making, text emerges as one of the most prevalent and valuable sources of information. From healthcare to public administration, vast amounts of knowledge are stored in textual form, fueling advances in research and practical applications. However, the inherent richness of textual data often encompasses personal and sensitive information, which, if improperly handled, could result in privacy leakages. Regulatory frameworks such as the European General Data Protection Regulation (GDPR) require anonymization of personal data before they can be released.

Anonymization precludes linkage between sensitive information and the subjects to whom the data refer, and thus brings the data outside the scope of the GDPR.

However, anonymization of textual data is a complex issue. Text data inherently carry nuanced, context-dependent information, making the identification and mitigation of disclosure risks particularly challenging [Lison et al., 2021, Csányi et al., 2021]. In the past, text anonymization has been mostly approached as a named entity recognition (NER) task, whereby named entities are detected and masked because, due to their specificity, they are assumed to encompass the most privacy-sensitive information. However, several studies [Lison et al., 2021, Hassan et al., 2023, Sánchez and Batet, 2016] have shown that NER-based methods offer weak protection due to the limited number of named entity types that they support.

On the other hand, Large Language Models (LLMs) [Brown et al., 2020, OpenAI, 2023, Touvron et al., 2023, Reid et al., 2024] have recently emerged as powerful tools for text processing and generation. They exhibit advanced proficiency in tasks like natural language understanding and automated reasoning, which makes them attractive for addressing the challenges associated with text anonymization. Early studies in this line [Bubeck et al., 2023, Patsakis and Lykousas, 2023, Singhal et al., 2024] suggest that LLMs might offer better protection than classic NER-based methods. However, due to the lack of a common and objective evaluation framework, the extent to which LLM-based text anonymization outperforms existing methods, as well as the influence of prompt variations on the protected outcomes, remains unclear.

1.1 Contributions

This paper presents the following contributions.

- We perform a critical analysis of the strengths and limitations of LLM-based anonymization methods.
- We propose an improved LLM-based anonymization through a refined prompt design. Empirical results shows that our prompt achieves the best balance between privacy protection and utility preservation across all analyzed methods, the best privacy protection in absolute terms, and a performance that scales proportionally with the capabilities of the underlying LLM.
- We propose a general and automatic evaluation framework for text anonymization, which does not rely on manual annotations or ground truth data.
- We apply our evaluation framework to a variety of text anonymization methods, including state-of-the-art approaches based on LLMs and classical mechanisms based on NER. As a result, we report clear and comparable performance figures that highlight the benefits and drawbacks of the different approaches.
- We also compare automatic anonymization methods with manual annotations, which have been considered as the gold standard for text anonymization so far. We find that LLM-based approaches significantly outperform them.

The rest of the document is organized as follows. Section 2 provides background on LLMs, text anonymization and its evaluation. Section 3 analyzes the current state-of-the-art in LLM-based methods and presents the proposed anonymization prompt. Section 4 describes our evaluation framework. Experimental details and evaluation results are reported in Section 5. The final Section 6 provides the conclusions and some lines of future research.

2 Background

2.1 Large Language Models

Large language models (LLMs) are large neural networks with billions to trillions of parameters, typically built using the Transformer architecture [Vaswani et al., 2017]. These models generate sequences by predicting the next token – representing a word or part of a word– one step at a time, in an autoregressive manner. They are first trained on vast and diverse corpus in a process called pre-training, enabling them to learn complex linguistic patterns, acquire broad general knowledge. Afterwards, they are fine-tuned to understand how to follow human-made instructions. Once pre-trained and fine-tuned, LLMs can be guided through natural language prompts to perform specific tasks. This approach has proven highly effective, often matching or even surpassing human performance across a wide range of applications [Bubeck et al., 2023, Tan et al., 2024, Gilardi et al., 2023].

Interacting with a LLM typically involves a conversational format, alternating between the roles of *User* (the human participant) and *Assistant* (the model). Optionally, an initial predefined *System* prompt can be included to set the overall behavior or tone of the model. For automatic task solving, the standard approach is to start with a prompt describing the task and providing the task data (with User and/or System role/s) and then use the Assistant response as solution. These prompts may include examples of correct solutions, which help the model to understand what it is expected to do. Depending on the number of examples provided, prompts/tasks are categorized as *zero-shot* (no examples), *one-shot* (one example), or *few-shot* (a small number of examples). Both LLM capabilities and prompt quality have a significant influence on response accuracy.

2.2 Text anonymization methods

Anonymization aims to prevent the released data from being linked to the subjects to whom the data are referred. To this end, it is necessary to mask identifying and quasi-identifying attributes [Hundepool et al., 2012], often referred to as personal identifiable information (PII). Whereas identifiers (such as full names or IDs) uniquely distinguish individuals and thus directly reveal their identities, quasi-identifiers are not unique for any individual, but can be by combination with other quasi-identifiers (*e.g.*, birth date + zip code). Text

anonymization involves suppressing identifiers and either suppressing or generalizing quasi-identifiers. A related concept is de-identification, which is a relaxed form of anonymization that mainly focuses on suppressing identifiers.

In textual data, any text span (*i.e.*, individual words or set of words) may act as a (quasi-)identifier. This, combined with the inherent complexity of natural language, makes PII detection a major challenge for text anonymization [Lison et al., 2021].

Most text anonymization methods rely on NER, by which the named entities detected in the text are detected and suppressed or replaced by their type (*e.g.*, PERSON, LOCATION...) [Johnson et al., 2020, Meystre et al., 2010, Yang and Garibaldi, 2015, Hassan et al., 2018, Neamatullah et al., 2008, Yogarajan et al., 2018, Huang et al., 2020, Mamede et al., 2016]. However, NER-based methods rely on the fundamentally flawed assumption that all PIIs are named entities [Lison et al., 2021, Hassan et al., 2023, Sánchez and Batet, 2016], whereas (quasi-)identifiers are not type-bounded. Furthermore, not all the named entities in a text are (quasi-)identifiers, because they might not refer to the subject to be protected, thereby leading to unnecessary masking that impairs the utility of the anonymized text.

In contrast, methods within the privacy-preserving data publishing (PPDP) field consider every text span as a potential PII [Hassan et al., 2023, Papadopoulou et al., 2022, Sánchez and Batet, 2016, Mosallanezhad et al., 2019, Staddon et al., 2007, Fernandes et al., 2019]. This intrinsically offers better protection than NER-based methods, at the cost of a higher complexity and, sometimes, worse utility preservation. PPDP-oriented methods often require extensive resources (*e.g.*, knowledge graphs encompassing all the concepts within the documents), produce an output that does not preserve the document's structure (*e.g.*, bag of words, word distributions), or suffer from scalability issues. Therefore, most of them are not applicable to real-world scenarios [Papadopoulou et al., 2022, Sánchez and Batet, 2016, Mosallanezhad et al., 2019, Staddon et al., 2007, Fernandes et al., 2019].

The advent of LLMs has recently sparked interest in their potential for text anonymization [Bubeck et al., 2023, Patsakis and Lykousas, 2023, Singhal et al., 2024]. As we detail in Section 2.1, in these approaches, LLMs are prompted for PII detection. Depending on the prompt design, a more NER-based or PPDP-oriented anonymization is obtained. Preliminary results suggest that the extensive knowledge of pre-trained LLMs effectively provides more nuanced and context-aware PII detection than traditional approaches. Nevertheless, the experimental evaluation reported so far [Bubeck et al., 2023, Patsakis and Lykousas, 2023, Singhal et al., 2024] lacks standardization due to being done ad hoc and with significantly different evaluation data, metrics, and models. Moreover, no comprehensive comparison of LLM-based methods has been reported so far. Therefore, there is no solid evidence on how much better LLM-based anonymization is, and how different prompts affect the protected outcome.

2.3 Evaluation

Data anonymization involves a trade-off: enhancing privacy reduces the analytical utility of the protected outcomes. Anonymization methods aim to optimize this trade-off, often seeking to maximize utility for a given privacy level. Thus, accurate privacy and utility evaluation is required for data holders to make an informed decision on the method to be used for their data release.

On the one hand, anonymization should prevent re-identification, meaning that the protected data cannot be unequivocally associated with a subject's identity. Thus, *privacy evaluation metrics* should measure the probability of re-identification. In tabular data, privacy models such as k-anonymity[Samarati, 2001] guarantee an *ex ante* upper bound for re-identification probability of $\frac{1}{k}$. For methods without formal privacy guarantees, an empirical evaluation of the residual re-identification risk is needed. For tabular data, the record linkage re-identification attack is the most widespread measure of risk [Domingo-Ferrer and Torra, 2003, Nin Guerrero et al., 2007, Torra et al., 2006, Torra and Stokes, 2012, Abril et al., 2012]. Record linkage aims to link quasi-identifiers in the anonymized dataset to identified public sources, which represent the background knowledge available to attackers. Re-identification risk is quantified as the ratio of correctly linked records.

Data utility preservation, on the other hand, is evaluated in tabular data by accounting for differences (or errors) between the original and anonymized data [Hundepool et al., 2012].

In contrast to the formal and objective evaluation of tabular data, the evaluation of text anonymization has been carried out primarily by comparing protected outcomes to human annotations [Johnson et al., 2020, Meystre et al., 2010, Yang and Garibaldi, 2015, Hassan et al., 2018, Neamatullah et al., 2008, Yogarajan et al., 2018, Huang et al., 2020, Mamede et al., 2016]. In this context, privacy protection and utility preservation are indirectly measured by means of recall and precision information retrieval metrics. On the one hand, recall quantifies the percentage of PIIs annotated by humans that the anonymization method has masked, and it is used as a proxy of the level of privacy protection attained. On the other hand, precision assesses the proportion of PIIs masked by the method that were also labeled as PIIs by humans. Because masked text spans that were not annotated by humans are deemed unnecessary, precision is considered as a measure of utility preservation. However, these evaluation means are severely limited [Lison et al., 2021, Manzanares-Salor et al., 2024]. On the one hand, human annotations are inherently imperfect, as we humans are prone to errors and biases. Empirical evidence shows that such inaccuracies can result in non-negligible re-identification risks [Manzanares-Salor et al., 2024]. In addition, there is no such thing as a single "ideal" anonymization, because several masking combinations of quasi-identifiers may be equally valid to prevent re-identification. Furthermore, precision and recall do not account for the significantly different influence that each (missed or overmasked) term might have on the re-identification risk or on data utility. Last but not least, manual annotations are costly to produce and require the input of several human experts.

Recently, some works [Patsakis and Lykousas, 2023, Nyffenegger et al., 2023] have proposed automatic privacy evaluations that use the LLMs they leveraged for anonymization to re-identify the individuals from the protected documents. While this approach directly targets the evaluation of re-identification risk, it faces a significant limitation: using the same tool for both anonymization and evaluation (i.e., the LLM) introduces biases. For instance, (quasi-)identifiers missed during anonymization might also be ignored during evaluation, leading to an underestimation of re-identification risk. On the other hand, utility evaluation has been largely ignored by these works.

3 Analysis and enhancement of LLM-based anonymization methods

This section examines the state-of-the-art in LLM-based PII detection. Building on this analysis, we introduce a novel approach that aims to improve current methods.

3.1 Analysis

In line with the discussion in Section 2.1, all LLM-based approaches rely on crafting specific prompts to direct a pre-trained LLM in detecting sensitive terms. In the following, we list and analyze the prompts proposed in the literature. Note that the "`[DOC_TO_PROTECT]`" mark will be replaced by the document to protect in the experiments. Moreover, for each sensitive term detected by the LLM, all its instances within the document are masked by replacing them with the tag "SENSITIVE" to generate the anonymized text.

- **Man vs the machine** [Patsakis and Lykousas, 2023]: In this work, the authors propose a short zero-shot prompt that uses both *System* and *User* roles. It offers a PPDP-oriented (rather than NER-oriented) PII detection by asking the LLM to detect every token potentially connected to the person to protect. This is equivalent to detect all the individual's attributes. Given the inherent uncertainty about which attributes may serve as (quasi-)identifiers, the choice of of treating all attributes as PII is deemed privacy-strict, as it prioritizes privacy to utility. The prompt also instructs the model to structure the response as a JSON-formatted list (each element is a detected sensitive term), which greatly simplifies response parsing.

> **System**: You are an efficient assistant. Keep your responses short and provide no explanations.
> **User**: Identify every token (including words, dates, and numerals) in the provided person description that can be connected to the person being talked about, such that if these tokens are removed the resulting text is anonymised. Return these words as a JSON-formatted list. The description of this person is: `[DOC_TO_PROTECT]`

– **Sparks** [Bubeck et al., 2023]: This study aims to explore the capabilities of pre-trained GPT-4 [OpenAI, 2023] across various tasks, including PII detection. To this end, the authors propose a zero-shot NER-oriented prompt, describing the 8 PII types to identify, along with some examples of them. While the approach covers a broad range of PII types, it likely inherits the privacy-related limitations standard of NER-based methods. The authors clearly denote the text to protect with the word "STATEMENT". For the response, they instruct the model to "*List these personally identifiable information as a python list using the format 'LIST:'.*" and "*Count the number of personally identifiable information in LIST and present the outcome using the format 'COUNT:'.*". This non-detailed response format often results in inconsistencies, such as lists with varying delimiters (e.g., hyphens, brackets, or line breaks) and additional details like PII types or explanations. These variations complicate parsing due to the lack of uniformity. To address this, we adopted a streamlined solution inspired by [Patsakis and Lykousas, 2023], which is to replace the original instructions with a single directive for a "*JSON-formatted list of strings*". This subtle adjustment greatly simplifies response parsing and ensures a fairer comparison by eliminating the need for complex and error-prone parsing code. The final prompt is:

User: INSTRUCTIONS:
1. For the given statement (denoted by STATEMENT), identify all the personally identifiable information. These include:
a. Names of people, including nicknames/aliases, usernames and initials.
b. Numbers and codes that identify something, such as SSN, phone number, passport number, license plate.
c. Places and locations, such as: cities, areas, countries, etc. or named infrastructures (bus stops, bridges, etc.)
d. Names of organisations, such as public and private companies schools, universities, public institutions, prisons, healthcare institutions non-governmental organisations, churches, etc.
e. Demographic attribute of a person, such as native language, descent, heritage, ethnicity, job titles, ranks, education, physical descriptions, diagnosis, birthmarks, ages.
f. Description of a specific date (*e.g.*, October 3, 2018), time (*e.g.*, 9:48 AM) or duration (*e.g.*, 18 years).
g. Description of a meaningful quantity, *e.g.*, percentages and/or monetary values.
h. Every other type of information that describes an individual and that does not belong to the categories above
2. List these personally identifiable information as a JSON-formatted list of strings.

TASK:

STATEMENT = "`[DOC_TO_PROTECT]`"

– **Student de-identification** [Singhal et al., 2024]: The authors of this work focus on de-identifying forum posts from students enrolled in massive open online courses. They frame it as a NER task by using a one-shot prompt that enumerates the considered PII types (without descriptions) and a example of a post. The example illustrates the (low) protection required for de-identification compared to the stricter requirements of anonymization (discussed in Section 2.2), as only the hyperlink is detected while course and company names are neglected. By focusing on de-identification and aligning with a NER-based approach, this prompt is expected to yield less private (but more utility-preserving) results, particularly for documents other than forum posts. This serves as a baseline to assess LLM capabilities under relaxed requirements. For the response format, the request is to "*Replace all removed PII with '[REDACTED]'. Ensure that the rest of the text remains unchanged, word for word.*". Similarly to the original *Sparks* prompt [Bubeck et al., 2023], this approach requires complex and error-prone parsing. Consequently, as with the previous prompt, we slightly modify it to request a JSON-formatted list, also adapting the example. The final prompt is:

User: Please return a JSON-formatted list with any personally identifiable information (PII) of the provided text. This includes names, company names, places of origin, current living locations, addresses, and social media links. If the text does not contain any PII, return an empty list.

For example, if the input is:
Dear All, You'll find below a Linkedin group I just create http://www.linkedin.com/groups?home=12345&abcd=789hm . It will be hold in English (as a second language) and is open to any student of Coursera for the 'Introduction to Financial Accounting'. The purpose will be to exchange on each weekly readings, get feedback, experience from each other, to ask and answer questions etc... Link you soon! Let's team work!

The output must be:
`["http://www.linkedin.com/groups?home=12345&abcd=789hm"]`

Please repeat this process with the following text:
`[DOC_TO_PROTECT]`

3.2 Our prompt

Inspired by the aforementioned LLM-based approaches, we propose a new prompt designed to enhance their effectiveness by strategically integrating what we perceive to be their core strengths. Our prompt is based on the hypothesis that these strengths can create synergies with each other, leading to measurable improvements over the current state-of-the-art. Moreover, we mitigated some limitations in the current LLM-based anonymization methods.

The core strengths of the different methods we aim to leverage are:

- **One-shot**: LLMs can generate reasonable responses with zero-shot prompts, such as those found in *Man vs the machine* or *Sparks*. However, using a one-shot prompt, like the one from *Student de-identification*, often provides the LLM a better understanding of the expected task. This approach makes it possible to provide a fine-grained sample of the required protection, beyond achieving anonymization. A well-crafted and comprehensive example is essential to ensure that the desired guidance is communicated effectively, minimizing the risk of the LLM deriving incorrect conclusions. For example, from the *Student de-identification* prompt, the model might infer that only hyperlinks are really significant, which may not align with the intended guidance. Consequently, we defined an example for our prompt in line with the aforementioned plan of detecting all individual's attributes.
- **Beyond NER limitations**: NER-based approaches fall short in privacy protection as they are confined to named entities (see Section 2.2). Despite the proficiency of LLMs in NER tasks, this limitation makes NER unsuitable

for comprehensive text anonymization. Instead, the *Man vs the machine* prompt introduces a broader strategy, aiming to detect (and subsequently mask) all attributes of the individual to protect. Among these attributes, there are the (quasi-)identifiers, so masking all of them should result in anonymity. For our prompt, we consider it preferable to detect and mask all attributes, which offers a robuster protection than NER-oriented approaches.
- **JSON-formatted response**: The output must adhere to a well-structured output format to ensure accurate and reliable parsing of detected PII. The *Man vs the machine* prompt proposes using a JSON-formatted list, with each element being a detected PII. This approach minimizes parsing errors due to its standardized nature. Moreover, LLMs' APIs such as OpenAI API allow to force the output to be a valid JSON, streamlining the process. We will follow this approach for our prompt.

Other secondary strengths we considered are the use of the *System* role from the *Man vs the machine* prompt, and the clear notation of the document to protect (*e.g.*, "STATEMENT") from the *Sparks* prompt.

Moreover, we addressed the following limitations of the current LLM-based anonymization methods:

- **Biased to NER**: NER-based anonymization is the predominant approach to text anonymization (see Section 2.2). Consequently, it is reasonable to expect that the vast majority of examples of text anonymization available in the internet will be NER-oriented. Since pre-trained LLMs derive much of their knowledge from internet data, they are likely influenced by this bias, even if not instructed for NER-oriented anonymization. To mitigate this issue, our prompt explicitly instructs the model to consider *all* types of terms, rather than being limited to named entity types.
- **Neglecting occurrences**: An in-depth analysis of anonymizations generated by the *Man vs the machine* prompt revealed occasional inconsistencies in the detection of different instances of the same sensitive attribute. For example, detecting the term "training", but overlooking the term "coach", even though both have a very similar meaning. If the link with "training" is a (quasi-)identifying attribute, such omissions will compromise anonymization, since any missed term related to "training" will allow the attacker to exploit the (quasi-)identifier. This issue likely arises because the LLM assumes that identifying one term implicitly covers its equivalents. To address this problem, our prompt emphasizes the need to detect all occurrences of sensitive terms.

By integrating all the discussed aspects together, we propose the following prompt:

System: You are an expert assistant. Provide no explanations and make your answers as long as required.
User: Identify every text span (including words, dates, and numerals) in the provided TEXT that defines an attribute of the corresponding person. Return them as a JSON-formatted list, with the verbatim text spans of TEXT (including formatting) corresponding to attributes. Ensure to capture all occurrences of these text spans, including synonyms, analogues, variations and specificities. Don't be limited to named entities, any text span can be an attribute.

TEXT:
It is believed that John Oldman was better as a coach than as an athlete. In fact, many people think Smith would not have made it as far as he did at the 2004 Olympics without Oldman's training. Oldman's disappearance in 2007 remains a mystery.
Assistant:`["John Oldman", "coach", "athlete", "Smith", "2004", "Olympics", "Oldman's", "training", "disappearance", "2007", "remains a mystery"]`
User: TEXT:
`[DOC_TO_PROTECT]`

4 Evaluation framework

On the basis discussed in Section 2.3, there is a need for a text anonymization evaluation framework directly grounded in privacy and utility principles that enables unbiased automatic assessments. Such a framework should incorporate a nuanced understanding of the informational value of terms (fundamentally determined by their semantics), and the fundamentals of re-identification. Moreover, by disentangling the evaluation process from human annotations, it should provide a more principled, objective, and practical approach to assessing anonymization methods. In the following, we describe the design of such a framework.

4.1 Privacy metric

Privacy protection is inversely proportional to the likelihood of re-identifying individuals within anonymized data. This likelihood can be assessed empirically by simulating a re-identification attack. For structured databases, it is standard to perform a record linkage attack, with its accuracy serving as indicator of the re-identification risk (refer to Section 2.3).

For textual data, [Manzanares-Salor et al., 2024] introduces a text re-identification attack (TRIA), which can be seen as the text equivalent of record linkage. To the best of our knowledge, it is the first of its kind in this regard. Both aim to link protected data (*i.e.*, document or record) with the individual they refer to,

by leveraging the publicly available background knowledge that might be exploited by attackers. This background knowledge ideally encompasses a superset of the individuals referred in the protected data. For example, if the protected documents are medical records from the patients of a hospital, the background knowledge might include identified social media posts from the inhabitants of the city where the hospital is located.

Regarding linkage, the (numerical or categorical) similarity measures used for classical record linkage are insufficient due to text inherent complexity. Subsequently, TRIA relies on a pre-trained language model, which is further pre-trained and fine-tuned on the background knowledge. The model acts as the linker, by predicting the individual from the background knowledge to whom most likely the (protected) document refers.

A successful link occurs when the individual for a protected document is correct predicted, thereby re-identifying the document's content. Therefore, the accuracy of TRIA for a set of protected documents effectively measures the text re-identification risk (TRIR):

$$TRIR = \frac{\#Successful\ links}{\#Links} \tag{1}$$

4.2 Utility metric

The utility of anonymized documents depends on how closely they resemble the original data (see Section 2.3). Applying this notion to textual data, we propose a utility metric that quantifies the amount of the original document's information that has been preserved in the anonymized outcome.

In computational linguistics, the informativeness of a term –referred to as its *information content* (IC)– can be derived by using Shannon's information theory[Shannon, 1948]:

$$IC(t) = -log(prob(t)) \tag{2}$$

As a result, highly probable (or predictable) terms will deemed less informative, while less probable terms will provide more information. The details of the probability calculation are outlined below.

The IC of a whole document can be defined as the sum of the IC from all its terms. In the case of protected documents, the IC of protected terms is excluded because text anonymization methods typically replace sensitive terms with generic placeholders (*e.g.*, "PERSON", "LOCATION", "***") that do not retain the original meaning. Based on this principle, we define *text information content* (TIC) as follows:

$$TIC(D) = \sum_{t \in D} IsNotProtected(t) \cdot IC(t) \tag{3}$$

where $IsNotProtected(t)$ is a binary function that returns 0 if the term t in document D is protected, and 1 otherwise.

To assess the proportion of information from the original document that is preserved in the protected document, we define *text preserved information* (TPI) as the ratio of their respective TICs:

$$TPI(P, O) = \frac{TIC(P)}{TIC(O)} \tag{4}$$

This ratio provides an intuitive measure of utility preservation.

To calculate the probabilities in Eq. 2, several methods have been proposed in the NLP literature. These include leveraging word frequencies within a corpus [Resnik, 1995] (even using the Web as a corpus [Sánchez et al., 2013, Sánchez and Batet, 2016]) analyzing the number of hypernyms and/or hyponyms in an ontology [Seco et al., 2004, Batet and Sánchez, 2020], and assessing predictability with a language model (LM) [Pilán et al., 2022]. The latter leverages the general language understanding capabilities of LMs, to predict a term based on its context. In this method, the context surrounding the target term (*e.g.*, the surrounding sentence) is provided to the model, with the target term replaced by the [MASK] token. The model then estimates the probability of the target term based on this context. For multi-token text spans, the minimum probability among the target term's tokens is selected. Notably, this method is the only one to consider compositionality, which asserts that the meaning of an expression arises from the meanings of its components and their combination rules. Additionally, it is the only technique capable of estimating probabilities for arbitrary terms, regardless of their presence in a specific corpus or ontology.

Given these advantages, we employ this approach to compute the IC of text spans. However, it was designed to estimate the IC for only a subset of terms, while Eq. 3 requires computing the IC for all text spans in a document. Directly replacing all spans with [MASK] would remove the necessary context for accurate predictions. To overcome this, we propose an alternating [MASK] substitution, by applying it to one out of every N terms while the rest provide context. Multiple prediction passes shift the [MASK] placement. For example, using $N = 2$, two passes are made: one masking odd terms and the other masking even terms. Larger N values improve the quality of probability estimates (as more context is considered) but increase computation time. The concrete implementation details employed in the reported evaluations are given in Section 5.2.

5 Evaluation results

In this section we provide a detailed description of the experimental setup, and we report, compare and discuss the evaluation results under our proposed framework. Our evaluation encompasses a wide variety of anonymization methods, together with an ablation study of our prompt (see Section 3.2).

5.1 Dataset

As common evaluation data, we leveraged the bibliographic abstracts presented in [Papadopoulou et al., 2022], which were extracted from 553 randomly selected

Wikipedia biographies. Wikipedia biographies are commonly employed as evaluation data for text anonymization due to their high density of (quasi-)identifying terms [Sánchez and Batet, 2016, Staddon et al., 2007, Sánchez et al., 2013, Hassan et al., 2023, Lison et al., 2021]. Moreover, biographies in this dataset predominantly feature relatively unknown individuals, such as World War II generals or lower-division football players. This makes the subjects in question unlikely to be familiar to pretrained LLMs, thereby presenting a more realistic scenario in which the subjects to be protected have not been significantly exposed to the LLM employed for anonymization.

Moreover, as far as we know, this is the largest domain-independent dataset for text anonymization that includes manual annotations, which are useful to have a human-based baseline for comparison. Annotations in this dataset were made by several human annotators (*i.e.*, one NLP researcher and four undergraduate students in law), and were based on the annotation guidelines described in [Pilán et al., 2022]. The process involved identifying sensitive terms, classifying them into semantic categories (such as person, date, organization), and deciding whether masking was necessary to protect the subjects' identities. The starting point were NER annotations obtained with *spaCy NER*[1], but human annotators were instructed to make any needed adjustments. The anonymization goal set for annotators was to mask enough text to render the subject referred in the text non-identifiable.

5.2 Implementation details

To enforce our privacy metric (TRIR), we scraped the Wikipedia to obtain the bodies corresponding to the 553 abstracts in the evaluation dataset, which we used as background knowledge for the re-identification attack. This defines a strong attacker, because background knowledge exactly matches the population of individuals in the protected dataset. The implementation employs the code available at `https://github.com/BenetManzanaresSalor/TextRe-Identification`, with all settings set to default values.

For our utility metric (TPI), we employ a pre-trained BERT language model (bert-base-uncased[2]) to calculate the IC of terms. To determine the appropriate value for N, we tested multiple increasing values, aiming to balance IC accuracy and runtime (see Section 4.2). Results varied with increasing N up to $N = 6$, after which further increments produced negligible differences while significantly increasing runtime. Based on these findings, we set $N = 6$ for TPI computation.

Regarding the LLMs employed for anonymization, it is expected that data holders will use the most advanced LLM that fits within their resource constraints. This is because the capabilities of LLMs are expected to have a major impact on response quality (see Section 2.1). Consequently, we selected a set of top-tier LLMs with different resource demands. When choosing these models, we recognized the pivotal role of OpenAI[3] in shaping the rapidly evolving field

[1] `https://spacy.io/api/entityrecognizer`
[2] `https://huggingface.co/google-bert/bert-base-uncased`
[3] `https://openai.com/`

of LLM technology. OpenAI often set the trend or the peak to be surpassed, consistently achieving the top rankings on the Chatbot Arena [Chiang et al., 2024][4], with larger and more sophisticated architectures. In fact, all LLM-based methods we analyzed in this work leveraged OpenAI models in their original papers [Bubeck et al., 2023, Patsakis and Lykousas, 2023, Singhal et al., 2024]. Subsequently, we selected three increasingly capable OpenAI models for our experiments:

- **GPT-3.5-Turbo-0125**[5]: GPT-3.5-turbo is an evolution from the original GPT-3[Brown et al., 2020] presented in 2022, and it is currently only accessible through an API. It was conceived as a fast, inexpensive model for simpler tasks. In Chatbot Arena, GPT-3.5-turbo is currently placed at the bottom of the top 100, with capabilities slightly better than *Meta-Llama-3.2-3B-Instruct*[6] or *Llama-2-70B-chat*[Touvron et al., 2023].
- **GPT-4o-mini-2024-07-18**[7]: GPT-4o-mini is the cost-efficient evolution of GPT-4 [OpenAI, 2023]. It is presented as an affordable and intelligent small model for fast, lightweight tasks. It currently ranks near the bottom of the top 10 in Chatbot Arena, with capabilities comparable to those of *Gemini-1.5-Flash-002* [Reid et al., 2024] and *Llama-3.1-Nemotron-70B-Instruct*[8].
- **GPT-4o-2024-09-03**[9]: GPT-4o is the most advanced evolution of GPT-4[OpenAI, 2023], and it serves as ChatGPT's flagship high-performance model for complex, multi-step, general tasks. In Chatbot Arena, the model currently ranks at the top 3, with similar capabilities to those of *Gemini-2.0*[10].

We leveraged the OpenAI API Python implementation [11], and forced JSON as response format, with default values for the rest of the settings.

5.3 Non-LLM-based baselines

In addition to evaluate LLM-based methods, we also considered a variety of non-LLM-based approaches to serve as baselines. These comprise both NER-based and PPDP-oriented methods and tools:

- **spaCy NER**[12]: The spaCy library offers an entity recognizer module that has been employed for automatic text anonymization in previous works [Hassan et al., 2023, Papadopoulou et al., 2022, Manzanares-Salor et al., 2024]. It

[4] https://lmarena.ai/
[5] https://platform.openai.com/docs/models/gpt-3-5-turbo
[6] https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/
[7] https://platform.openai.com/docs/models/gpt-4o-mini
[8] https://huggingface.co/nvidia/Llama-3.1-Nemotron-70B-Instruct
[9] https://platform.openai.com/docs/models/gpt-4o
[10] https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/
[11] https://platform.openai.com/docs/api-reference/chat/create
[12] https://spacy.io/api/entityrecognizer

allows for the detection of the following named entity types: ORG, PERSON, DATE, TIME (*i.e.*, times smaller than a day), GPE (*e.g.*, cities, regions, countries), LOC (*i.e.*, non-GPE locations such as mountain ranges), LANGUAGE, EVENT (*e.g.*, wars), NORP (*i.e.*, nationalities or religious or political group), WORK_OF_ART (*e.g.*, songs, films), PRODUCT (*e.g.*, smartphone), FAC (*e.g.*, buildings), LAW (*i.e.*, named documents made into laws), ORDINAL, PERCENT, CARDINAL, QUANTITY and MONEY. In particular, we leveraged the en_core_web_lg neural model trained on the Ontonotes v5 corpus [Weischedel et al., 2011].

- **Microsoft Presidio**[13]: This library is specifically designed for text anonymization. It is based on NER and supports the following sensitive entity types: PERSON, DATE_TIME, NRP (*i.e.*, person's nationality, religious or political group) and LOCATION.
- **Disclosure protection based on word embeddings** [Hassan et al., 2023]: This PPDP-oriented method uses a Word2Vec [Mikolov et al., 2013] model trained from scratch on a corpus that includes the documents to be protected and, optionally, external documents that provide general in-domain knowledge. The word embedding model is designed to identify text spans that semantically resemble individuals requiring protection. Since word embeddings capture the semantic relationships between words, terms more likely contributing to re-identification are expected to appear close to an individual's name within the embedding space, because they will have a tendency to co-occur. Words with a cosine similarity to the individual's name exceeding a user-defined threshold are considered as (quasi-)identifiers, and are subjected to masking. We trained the Word2Vec model on the 553 biographies to be protected (including both abstracts and bodies), alongside a corpus of 19,000 Wikipedia biographies, as proposed in the original paper [Hassan et al., 2023]. We employed the same similarity threshold (0.25) used in the original paper's experiments (which were also conducted on Wikipedia biographies), and the public implementation available at `https://github.com/fadiabdulf/automatic_text_anonymization`. In the remainder, this method will be referred to as *Word2Vec*. Requiring only a basic word embedding model and (optionally) a set of documents from a domain similar to those to be protected, this is probably one of the most practical PPDP-oriented methods.
- **k-anonymity-based on knowledge graphs** [Papadopoulou et al., 2022]: A PPDP-oriented approach that leverages a knowledge graph derived from Wikidata's term distribution[14] to enforce k-anonymity on a set of documents. The method identifies terms that appear in fewer than k documents as sensitive, and iteratively masks them until all combinations of terms occur in at least k documents. Term masking can be done by using either a *greedy* approach, which masks the least frequent terms first, or a *random* approach, which arbitrarily selects sensitive terms for masking. The greedy technique

[13] `https://github.com/microsoft/presidio`
[14] `https://www.wikidata.org/`

is observed to mask fewer terms overall, which may better preserve the documents' utility. We selected this technique and refer to it as *k-anonymity Greedy*. Since the authors tested the methods on the same corpus of 553 abstracts used here, we directly employed the results they provided results (with $k = 5$). Note that, by requiring a knowledge graph containing the terms of the documents to be protected, this PPDP-oriented approach cannot be easily applied to datasets other than Wikipedia data.

In addition to the automatic approaches above, we also evaluated the manual annotations from the evaluation dataset described in Section 5.1.

The PPDP-oriented methods sensitive terms were masked by replacing them with the tag "SENSITIVE". For *spaCy NER*, *Microsoft Presidio*, and *manual annotations*, terms were masked by substituting them with their semantic category label (*e.g.*, PERSON, DATE), as it is standard in NER-oriented anonymization.

5.4 Results and discussion

As discussed in Section 2.3, anonymization methods should optimize the trade-off between privacy protection and utility preservation. A comprehensive evaluation thus requires the simultaneous analysis of both variables, proving that a method is irrefutably better only if it provides better privacy for the same utility or vice versa. On these basis, Figure 1 depicts a graph with vertical axis for our privacy metric (re-identification risk, *i.e.*, TRIR) and horizontal axis for utility preservation (*i.e.*, TPI). In this way, each anonymization method is represented as a point in this 2D space. The lower the TRIR and the higher the TPI, the better. Note also that a TRIR of 25% does not imply that 1 of each 4 documents can be unequivocally re-identified, but that 25% of the attacker guesses are correct, because the attacker has no way to know if a guess is correct or not. In terms of probabilistic k-anonymity [Samarati, 2001, Soria-Comas and Domingo-Ferrer, 2012], that would be equivalent to $k = 4$ ($\frac{1}{4} = 0.25$). LLM-based methods are represented by three connected points, corresponding to the three increasingly capable models listed in Section 5.2.

First, we can see that LLM-based methods obtain better or similar privacy protection for better or similar utility than any other approach. In other words, prompting pre-trained LLMs outperforms or matches current non-LLM-based approaches, including *manual annotations*. For example, *Student de-identification*, for an equivalent TRIR, obtains a TPI between 15% and 20% greater than NER-based methods. In the same way, *Sparks* obtains a TRIR between 5% and 12% lower than *spaCy NER* or *Manual annotations* with even slightly better TPI. The only exception to this Pareto optimality is *k-anonymity Greedy* that, even if no direct comparison is possible (no similar enough TRIR or TPI), it seems to provide the best utility for a $TRIR = 9\%$. However, it should be noted that this method requires a knowledge graph containing all the concepts of the corpus, which is hardly feasible outside Wikipedia-based datasets.

Secondly, prompting proved to be more influential than the actual LLM in use, with almost no overlap on the TRIR axis. When comparing prompts using

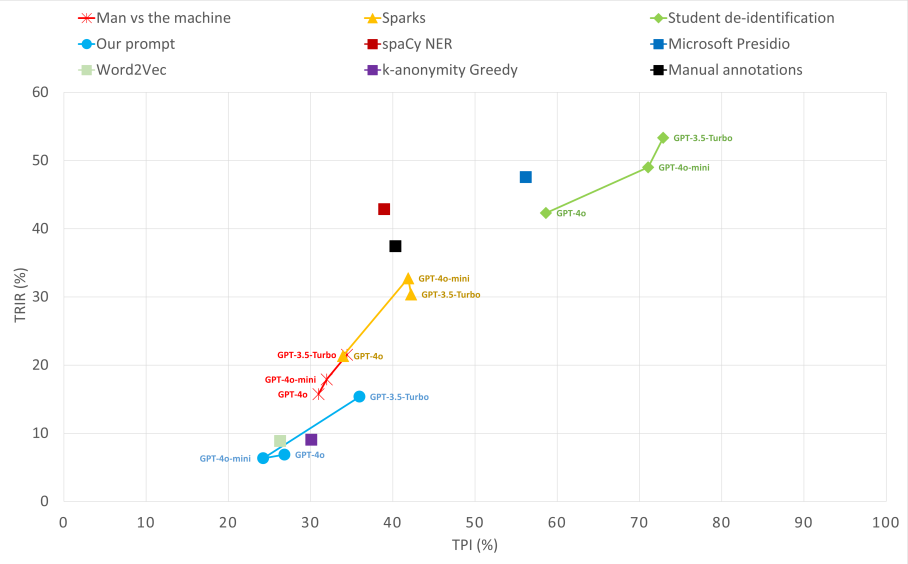


Fig. 1. Privacy and utility evaluation of the considered anonymization methods

the same LLM version, differences range between 5% and 40% for TRIR, and between 2% and 50% for TPI. It is common for a superior prompt using GPT-3.5-Turbo to obtain similar or even better results than a worse prompt leveraging the more advanced GPT-4o model.

In any case, *our prompt* obtained the best results, with better TPI for an equivalent TRIR and the lowest TRIR. The closest competitor prompt, *Man vs the machine*, obtained a similar TRIR (15%) by using GPT-4o than *our prompt* with GPT-3.5-Turbo, but with a 5% worse TPI. In other words, *our prompt* provides the best privacy-utility trade-off regardless the LLM version in use. This outstanding results support the reasoning in Section 3.2 about the strengths and limitations considered for the definition of *our prompt*. Table 1 illustrates this superiority by comparing the anonymized outcomes of all LLM-based methods. In particular, the other methods missed several highly disclosive information, such as the subject's profession, birth date, and/or nationality.

It is worth noting that, despite the fact that the usage of a more capable LLM tended to improve the results, the actual benefits were inconsistent between prompts. Ideally, improvement should be proportional to the capabilities of the model (*e.g.*, score increase in ChatBot Arena [Chiang et al., 2024]). Nonetheless, *Sparks* obtained better results with GPT-3.5-Turbo than with GPT-4o-mini, while GPT-4o produced the best results; besides, *Man vs the machine* and *our prompt* benefited from the LLM's capability increase, but improvement was more noticeable when moving from GPT-3.5-Turbo to GPT-4o-mini than from GPT-4o-mini to GPT-4o (inversely to *Sparks* and *Student de-identification*). This inconsistencies can be argued by the fact that LLM capabilities vary greatly

depending on the task (the general capability reported in ChatBot Arena is an average across many heterogeneous tasks), and may find it harder or easier to follow a particular prompt.

Interesting insights can be extracted from the analysis of the privacy protection achieved by the different approaches. On the weaker protection side (*i.e.*, TRIR greater than 35%) we find *Student de-identification*, NER-based methods (*i.e.*, *spaCy NER* and *Microsoft Presidio*) and the *manual annotations*. Even though de-identification and NER-based techniques were expected to provide little privacy protection (see Section 2.2), it is surprising that the *manual annotations* fell into the same group. Their proximity to *spaCy NER* (which was the starting point of annotations) suggests that humans did not consider that much additional masking was necessary. In addition, it is hypothesized that humans may miss some quasi-identifiers due to the inherently difficulty of considering many information simultaneously [Manzanares-Salor and Sánchez, 2025]. Moreover, *spaCy NER* and *manual annotations* do not standout on utility preservation, even when compared to other non-LLM-based approaches: their TPI values are only 10% greater than those of PPDP-based methods, which obtained TRIR figures more than 25% lower.

Looking at the methods offering the best protection (*i.e.*, TRIR lower than 35%), we have LLM-based approaches and the two PPDP-oriented techniques. We considered TRIR values of 9% or lower as robust protection[15], and those are only reached by the PPDP-based approaches and *our prompt*. Among these, the only practical solutions (*i.e.*, not requiring complex resources such as comprehensive knowledge graphs) are *our prompt* and *Word2Vec*. On the one hand, *our prompt* provides slightly better TPI for a TRIR slightly lower than *Word2Vec*. On the other hand, any LLM-based approach offers greater practicality, requiring only a pre-trained LLM and a prompt, wheres *Word2Vec* involves training a word embedding model from scratch by using the protected documents and, ideally, additional domain-specific texts.

5.5 Ablation study of our prompt

To assess the impact of each component of *our prompt* and confirm that each one contributes meaningfully to its performance, we defined the following ablated versions:

- **Excluding example**: Converts the task to a zero-shot setup by suppressing the text about John Oldman and the example of the *Assistant* response. That is, the ablated prompt uses only *System* and *User* roles.
- **Excluding no NER limitation**: Omits the instruction "*Don't be limited to named entities, any text span can be an attribute.*". As discussed in Section

[15] According to [El Emam, 2013], a strong degree of anonymity corresponds to a re-identification risk of approximately 9% or lower, (equivalent to $k = 11$ or greater in probabilistic k-anonymity terms [Samarati, 2001, Soria-Comas and Domingo-Ferrer, 2012]), as it has been explicitly endorsed by two independent health regulators [European Medicines Agency, 2014, Health Canada, 2019].

Table 1. Example of anonymized outcomes of LLM-based methods. Each PII detected by each method is replaced by the word *SENSITIVE*

.

Method	Text
Original text	Andrés Alejandro Palomeque González (July 1, 1971 – March 22, 2009) was a Mexican luchador (Spanish for "masked professional wrestler"). He is best known for appearing under the stage name Abismo Negro, which is Spanish for "Black Abyss", in the Asistencia Asesoría y Administración (AAA) promotion.
Man vs the machine	SENSITIVE SENSITIVE SENSITIVE SENSITIVE (SENSITIVE – SENSITIVE) was a SENSITIVE SENSITIVE (Spanish for "masked SENSITIVE wrestler"). He is best known for appearing under the stage name SENSITIVE, which is Spanish for "SENSITIVE", in the SENSITIVE (SENSITIVE) promotion.
Sparks	SENSITIVE (SENSITIVE – SENSITIVE) was a SENSITIVE luchador (Spanish for "masked professional wrestler"). He is best known for appearing under the stage name SENSITIVE, which is Spanish for "SENSITIVE", in the SENSITIVE promotion.
Student de-identification	SENSITIVE (July 1, 1971 – March 22, 2009) was a Mexican luchador (Spanish for "masked professional wrestler"). He is best known for appearing under the stage name SENSITIVE, which is Spanish for "Black Abyss", in the SENSITIVE promotion.
Our prompt	SENSITIVE (SENSITIVE – SENSITIVE) was a SENSITIVE SENSITIVE (Spanish for "SENSITIVE"). He is best known for appearing under the stage name SENSITIVE, which is Spanish for "SENSITIVE", in the SENSITIVE (SENSITIVE) promotion.

3.2, not explicitly asking to avoid this bias would likely produce results closer to (weaker) NER-based anonymization.

– **Excluding occurrences**: Omits the sentence "*Ensure to capture all occurrences of these text spans, including synonyms, analogues, variations and specificities*". As explained in Section 3.2, this instruction aims to prevent the model from missing alternative terms related to sensitive attributes (*e.g.*, masking " training" but not "coach"). Without this guidance, we

expect weaker anonymization due to leaving more terms exposed for re-identification.

Figure 2 depicts the results of the ablations above in comparison with the complete prompt on the same basis as for Figure 1.

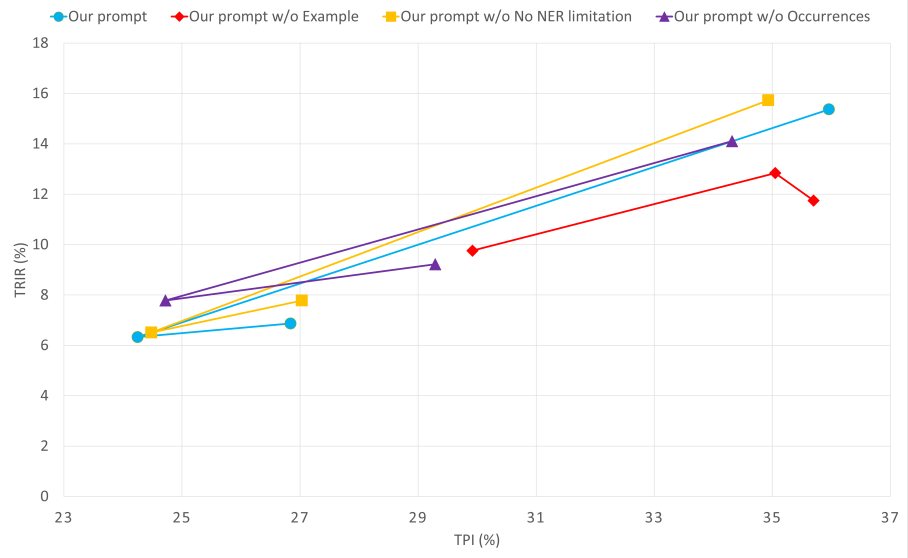


Fig. 2. Privacy and utility evaluation across ablations of our proposed prompt. Starting from the rightmost point of each line, the models used are GPT-3.5-Turbo, GPT-4o-mini and GPT-4o, respectively.

Results indicate that both the *no NER limitation* and *occurrences* ablations yielded similar or worse results, often producing a higher TRIR. In contrast, removing the example led to noticeably different outcomes. First, TRIR values are consistently lower for an equivalent TPI. Despite this could be considered an improvement, it also limited the lowest TRIR to 9% versus the 6% obtained by the non-ablated prompt. Secondly, this ablation altered the performance benefits across LLMs; as with *Sparks*, GPT-4o-mini's results were worse than those from GPT-3.5-Turbo.

We can thus conclude that each component tested is shown to contribute to privacy protection improvement (*i.e.*, decrease TRIR) and/or reaping the benefits of using a more capable LLM.

6 Conclusions

We have analyzed and evaluated under a common and objective framework the state-of-the-art in (LLM-based) text anonymization. Based on our results, we can draw the following conclusions:

- NER-based approaches, which have dominated text anonymization for the past two decades, are largely obsolete. These methods fail to capture the notion of risk effectively since they are not designed to understand what specific information (individuals, organizations, etc.) needs protection. As a result, they offer non-configurable privacy protection, which is likely insufficient for most data release scenarios.
- In contrast, LLM-based approaches demonstrate a better ability to assess and mitigate risks. These techniques are effective and straightforward to apply and require no external resources or complex configurations. Moreover, we have shown that prompts can be incrementally improved by leveraging the strengths observed in previous attempts. However, these methods are inherently limited in their ability to configure the balance between privacy and utility. Configuration comes from prompt engineering, a process that, while accessible, often yields unpredictable results due to the opaque nature of LLMs. They also lack mechanisms to provide *ex ante* guarantees regarding the level of privacy attained. As a result, tailoring or controlling the outcomes to align with desired anonymization requirements is mostly an empirical process.
- As it has been observed in other NLP-oriented tasks [Gilardi et al., 2023], manual annotations for text anonymization are not competitive against LLMs. This highlights the need to step away from manual methods, both for text anonymization and for privacy/utility evaluation.
- Given the suboptimal protection provided by manual annotations, using them as evaluation ground truth –as it has been done so far– is far from ideal. Instead, automated and purpose-built metrics to measure the actual residual re-identification risk or the observed analytical utility –an approach widely adopted in the SDC literature for structured databases over the past three decades– are far more desirable. These kind of metrics enable more objective, consistent, and reproducible evaluations of anonymization methods.

In conclusion, the field of text anonymization is evolving rapidly, with LLM-based methods paving the way for more effective and nuanced approaches. Nonetheless, significant challenges persist, such as the need for empirical experimentation and evaluation, and the lack of formal privacy guarantees on the protected outcomes.

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). The first author is also supported by the Spanish Government under an FPU grant (ref. FPU23/01785).

References

Abril et al., 2012. Abril, D., Navarro-Arribas, G., and Torra, V. (2012). Improving record linkage with supervised learning for disclosure risk assessment. *Information Fusion*, 13(4):274–284.

Batet and Sánchez, 2020. Batet, M. and Sánchez, D. (2020). Leveraging synonymy and polysemy to improve semantic similarity assessments based on intrinsic information content. *Artificial Intelligence Review*, 53(3):2023–2041.

Brown et al., 2020. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., and Askell, A. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Neural Information Processing Systems Foundation.

Bubeck et al., 2023. 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., and Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

Chiang et al., 2024. Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., and Stoica, I. (2024). Chatbot arena: An open platform for evaluating llms by human preference.

Csányi et al., 2021. Csányi, G. M., Nagy, D., Vági, R., Vadász, J. P., and Orosz, T. (2021). Challenges and open problems of legal document anonymization. *Symmetry*, 13(8):1490.

Domingo-Ferrer and Torra, 2003. Domingo-Ferrer, J. and Torra, V. (2003). Disclosure risk assessment in statistical microdata protection via advanced record linkage. *Statistics and Computing*, 13(4):343–354.

El Emam, 2013. El Emam, K. (2013). *Guide to the de-identification of personal health information*. CRC Press.

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

Fernandes et al., 2019. Fernandes, N., Dras, M., and McIver, A. (2019). Generalised differential privacy for text document processing. In *International Conference on Principles of Security and Trust*, pages 123–148. Springer.

Gilardi et al., 2023. Gilardi, F., Alizadeh, M., and Kubli, M. (2023). Chatgpt outperforms crowd-workers for text-annotation tasks. *CoRR*, abs/2303.15056.

Hassan et al., 2018. Hassan, F., Domingo-Ferrer, J., and Soria-Comas, J. (2018). Anonymization of unstructured data via named-entity recognition. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 296–305, Mallorca, Spain. Springer.

Hassan et al., 2023. Hassan, F., Sánchez, D., and Domingo-Ferrer, J. (2023). Utility-preserving privacy protection of textual documents via word embeddings. *IEEE Transactions on Knowledge and Data Engineering*, 35:1058–1071.

Health Canada, 2019. Health Canada (2019). Guidance document on public release of clinical information.

Huang et al., 2020. Huang, Y., Song, Z., Chen, D., Li, K., and Arora, S. (2020). Texthide: Tackling data privacy in language understanding tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1368–1382, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hundepool et al., 2012. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., and De Wolf, P.-P. (2012). *Statistical Disclosure Control*, volume 2. Wiley New York, New York.

Johnson et al., 2020. Johnson, A. E., Bulgarelli, L., and 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*, pages 214–221, Toronto, Ontario, Canada. Association for Computing Machinery.

Lison et al., 2021. Lison, P., Pilán, I., Sánchez, D., Batet, M., and 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)*, pages 4188–4203. Association for Computational Linguistics.

Mamede et al., 2016. Mamede, N., Baptista, J., and Dias, F. (2016). Automated anonymization of text documents. In *IEEE Congress on Evolutionary Computation*, pages 1287–1294, Vancouver, BC, Canada. IEEE.

Manzanares-Salor et al., 2024. Manzanares-Salor, B., Sánchez, D., and Lison, P. (2024). Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack. *Data Mining and Knowledge Discovery*, (In press).

Manzanares-Salor and Sánchez, 2025. Manzanares-Salor, B. and Sánchez, D. (2025). Enhancing text anonymization via re-identification risk-based explainability. *Knowledge-Based Systems*, 310:112945.

Meystre et al., 2010. Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., and 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):1–16.

Mikolov et al., 2013. Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781*.

Mosallanezhad et al., 2019. Mosallanezhad, A., Beigi, G., and Liu, H. (2019). Deep reinforcement learning-based text anonymization against private-attribute inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2360–2369, Hong Kong, China. Association for Computational Linguistics.

Neamatullah et al., 2008. Neamatullah, I., Douglass, M. M., Lehman, L.-W. H., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G., and Clifford, G. D. (2008). Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1):1–17.

Nin Guerrero et al., 2007. Nin Guerrero, J., Herranz Sotoca, J., and Torra i Reventós, V. (2007). On method-specific record linkage for risk assessment. In *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, pages 1–12. UNECE.

Nyffenegger et al., 2023. Nyffenegger, A., Stürmer, M., and Niklaus, J. (2023). Anonymity at risk? assessing re-identification capabilities of large language models. *CoRR*, abs/2308.11103.

OpenAI, 2023. OpenAI (2023). GPT-4 Technical Report. *arXiv e-prints*.

Papadopoulou et al., 2022. Papadopoulou, A., Lison, P., Øvrelid, L., and Pilán, I. (2022). Bootstrapping text anonymization models with distant supervision. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4477–4487, Marseille, France. European Language Resources Association.

Patsakis and Lykousas, 2023. Patsakis, C. and Lykousas, N. (2023). Man vs the machine in the struggle for effective text anonymisation in the age of large language models. *Scientific Reports*, 13(1):16026.

Pilán et al., 2022. Pilán, I., Lison, P., Ørelid, L., Papadopoulou, A., Sánchez, D., and Batet, M. (2022). The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(8):1053–1101.

Reid et al., 2024. Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T. P., Alayrac, J., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., Antonoglou, I., Anil, R., Borgeaud, S., Dai, A. M., Millican, K., Dyer, E., Glaese, M., Sottiaux, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Molloy, J., Chen, J., Isard, M., Barham, P., Hennigan, T., McIlroy, R., Johnson, M., Schalkwyk, J., Collins, E., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Meyer, C., Thornton, G., Yang, Z., Michalewski, H., Abbas, Z., Schucher, N., Anand, A., Ives, R., Keeling, J., Lenc, K., Haykal, S., Shakeri, S., Shyam, P., Chowdhery, A., Ring, R., Spencer, S., Sezener, E., and et al. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Resnik, 1995. 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*, IJCAI'95, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Samarati, 2001. Samarati, P. (2001). Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027.

Sánchez and Batet, 2016. Sánchez, D. and Batet, M. (2016). C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163.

Sánchez et al., 2013. Sánchez, D., Batet, M., and Viejo, A. (2013). Automatic general-purpose sanitization of textual documents. *IEEE Transactions on Information Forensics and Security*, 8(6):853–862.

Seco et al., 2004. Seco, N., Veale, T., and Hayes, J. (2004). An intrinsic information content metric for semantic similarity in wordnet. In *Proceedings of the 16th European Conference on Artificial Intelligence*, ECAI'04, pages 1089–1090, NLD. IOS Press.

Shannon, 1948. Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Singhal et al., 2024. Singhal, S., Zambrano, A. F., Pankiewicz, M., Liu, X., Porter, C., and Baker, R. S. (2024). De-identifying student personally identifying information with gpt-4. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 559–565.

Soria-Comas and Domingo-Ferrer, 2012. Soria-Comas, J. and Domingo-Ferrer, J. (2012). Probabilistic k-anonymity through microaggregation and data swapping. In *2012 IEEE International Conference on Fuzzy Systems*, pages 1–8.

Staddon et al., 2007. Staddon, J., Golle, P., and Zimny, B. (2007). Web-based inference detection. In *USENIX Security Symposium*, pages 1–16, Boston, MA, USA. Association for Computing Machinery.

Tan et al., 2024. Tan, E. J. L., Ramos, K. A. L., Nazario, M. E. K. B., Lim, S. V. D., and Chu, S. B. (2024). AI to the test: Measuring chatgpt's objective accuracy in the sats in comparison to human performance. In Shahriar, H., Ohsaki, H., Sharmin, M., Towey, D., Majumder, A. K. M. J. A., Hori, Y., Yang, J., Takemoto, M., Sakib, N., Banno, R., and Ahamed, S. I., editors, *48th IEEE Annual Computers, Software, and Applications Conference, COMPSAC 2024, Osaka, Japan, July 2-4, 2024*, pages 153–158. IEEE.

Torra et al., 2006. Torra, V., Abowd, J. M., and Domingo-Ferrer, J. (2006). Using mahalanobis distance-based record linkage for disclosure risk assessment. In *Privacy in Statistical Databases*, pages 233–242. Springer.

Torra and Stokes, 2012. Torra, V. and Stokes, K. (2012). A formalization of record linkage and its application to data protection. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 20(06):907–919.

Touvron et al., 2023. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Vaswani et al., 2017. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, l., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, Long Beach, CA, USA. Neural Information Processing Systems Foundation.

Weischedel et al., 2011. Weischedel, R., Hovy, E., Marcus, M., Palmer, M., Belvin, R., Pradhan, S., Ramshaw, L., and Xue, N. (2011). *OntoNotes: A large training corpus for enhanced processing*, book section 1, pages 54–63. Springer, New York.

Yang and Garibaldi, 2015. Yang, H. and Garibaldi, J. M. (2015). Automatic detection of protected health information from clinic narratives. *Journal of Biomedical Informatics*, 58:S30–S38.

Yogarajan et al., 2018. Yogarajan, V., Mayo, M., and Pfahringer, B. (2018). A survey of automatic de-identification of longitudinal clinical narratives. *arXiv:1810.06765*.