Protecting De-identified Documents from Search-based Linkage Attacks

Anonymous ACL submission

Abstract

While de-identification models can help conceal the identity of the individual(s) mentioned in a document, they fail to address *linkage risks*, defined as the potential to map the de-identified text back to its source. One straightforward way to perform such linkages is to extract phrases from the de-identified document and then check their presence in the original dataset.

This paper presents a method[1] to counter search-based linkage attacks while preserving the semantic integrity of the text. The method proceeds in two steps. We first construct an *inverted index* of the N-grams occurring in the text collection, making it possible to efficiently determine which N-grams appear in less than k documents (either alone or in combination with other N-grams). An LLM-based rewriter is then iteratively queried to reformulate those spans until linkage is no longer possible. Experimental results on two datasets (court cases and Wikipedia biographies) show that the rewriting method can effectively prevent search-based linkages while remaining faithful to the original content. However, we also highlight that linkages remain feasible with the help of more advanced, semantics-oriented approaches.

1 Introduction

Most text documents contain personal information in one form or another. To alleviate privacy concerns and abide to data protection regulations, many types of documents, such as administrative decisions, court cases or patient records, must be *de-identified* prior to transferring them to third parties or releasing them publicly. De-identification techniques typically operate by detecting text spans that express *personally identifiable information* (PII) and subsequently masking those from the document (Neamatullah et al., 2008; Dernoncourt et al., 2017; Lison et al., 2021; Szawerna et al., 2024; Staab et al., 2024; Deußer et al., 2025).

[1]The code will be published on GitHub after peer-review.

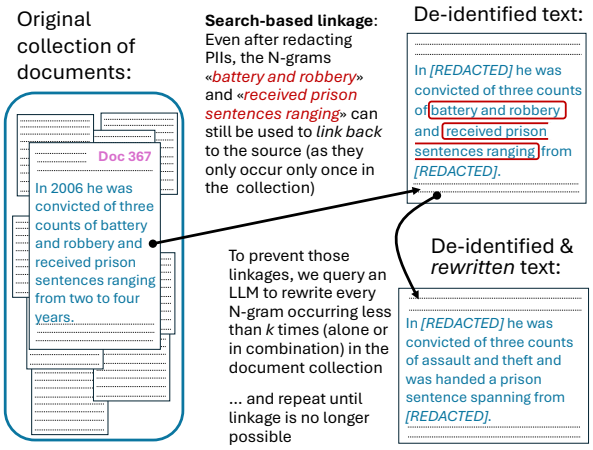

Figure 1: To prevent search-based linkages between a de-identified document and its original source, we first identify all N-grams occurring less than k times in the collection, and then use an instruction-tuned LLM to reformulate each of those N-grams (in their local context) until the linkages are averted.

De-identification techniques primarily focus on reducing the *re-identification risk* – that is, the risk that an adversary might infer the identity of the human individual(s) referred to in the document based on the PIIs occurring in the text. However, according to data protection regulations such as the European GDPR, data owners should also mitigate the risk of *record linkages*, which happens when the de-identified data remains linkable to other sources, and in particular to the original data collection. In Europe, this risk is notably emphasised in the Opinion of the Article 29 WP on Anonymisation Techniques (Article 29 Data Protection Working Party, 2014), which plays a key role in defining how anonymisation and pseudonymisation should be interpreted under the GDPR. The Opinion highlights the *(un)linkability* of pseudonymised data with its original source as one of the three criteria (along with the prevention of singling out and inference) that anonymisation techniques should meet.

Linkability can indeed allow sensitive information to be traced back to the original records, with potentially harmful consequences for the individuals in question, for instance if de-identified records labelled with health concerns are used by insurance firms to deny medical cover.

One simple but powerful technique for establishing such linkages is *phrase search* (Weitzenboeck et al., 2022). This method entails extracting phrases from the de-identified text and then searching for their occurrence in the original document collection. As many longer N-grams will only occur once in a given corpus, (Silva and Cunha, 2020), phrase search can efficiently connect the edited texts back to their source, as illustrated in Fig. 1.

Phrase search attacks can be conducted by adversaries with knowledge of only a few N-grams occurring in the original document. For instance, a de-identified patient record d' from a particular hospital may include the mention of a rare diagnosis named X, or include a particular sentence whose writing style reveals the identity of the doctor in charge of the patient. An adversary with the knowledge that an individual diagnosed with X was treated in that hospital (and presumably had X mentioned in their original patient record d) could then infer a link between this de-identified record d' and the original patient record d.

This paper presents a method for preventing such search-based linkages through an *inverted index* that captures the occurrence of N-grams across the collection. This inverted index is employed to efficiently derive the N-grams – and combinations of N-grams – that are associated with fewer than k documents and should therefore be rephrased when they occur in a given document. An instruction-tuned LLM is then queried with the text along with this list of N-grams to edit. This process is repeated until the text no longer contains any N-gram (or combination of N-grams) that could uniquely identify it within the original corpus.

Using two datasets (respectively based on international court cases and Wikipedia biographies), we show that existing text rewriting methods are poor at preventing search-based linkages. In contrast, the method proposed in this paper is able to drastically reduce linkage risks while remaining truthful to the original content of the document. However, more sophisticated linkage attacks remain possible for adversaries with access to the full collection of original documents.

The paper is structured as follows. Section 2 reviews related work on text de-identification, record linkages and paraphrasing. Section 3 describes the threat model along with the text rewriting method. This method is evaluated in Section 4 on two datasets. Finally, Section 5 discusses the results, followed by a conclusion in Section 6.

2 Related Work

Text de-identification The goal of text de-identification, sometimes also called *PII removal*, *text sanitisation* or *pseudonymisation*, is to detect spans expressing Personally Identifiable Information (PII) and subsequently mask those from the documents. Those PIIs may include both direct identifiers, such as full person names, phone numbers or home addresses, but also so-called *quasi*-identifiers, which do not disclose the person identity in isolation, but may do so when combined with other quasi-identifiers and background knowledge. Such quasi-identifiers encompass a wide range of information, such as year of birth, professional occupation, gender, hometown or nationality. Since both direct identifiers and quasi-identifiers occurring in a text may increase the risk of revealing the identities of the individuals mentioned, they should generally be masked or redacted.

Methods for text de-identification may rely on either heuristics or rule-based approaches (Neamatullah et al., 2008), information-theoretic measures (Sánchez and Batet, 2016), statistical or neural sequence labelling models (Dernoncourt et al., 2017; Eder et al., 2019; Pilán et al., 2022; Kleinberg et al., 2022; Hahm et al., 2025), unsupervised methods (Morris et al., 2022) and more recently LLMs, either simply prompted or fined-tuned to the task (Liu et al., 2023; Staab et al., 2024; Pissarra et al., 2024; Frikha et al., 2025).

Several studies have also proposed privacy-preserving text rewriting methods, notably to obfuscate demographic indicators (Xu et al., 2019), conceal author identity (Weggenmann et al., 2022; Bao and Carpuat, 2024), or implement differential privacy (DP) mechanisms (Igamberdiev and Habernal, 2023; Meisenbacher et al., 2024). However, those text rewriting techniques produce fully transformed texts that do not retain the document's original content. While such transformations are suitable for tasks such as synthetic data generation, they are inadequate for many downstream uses of text de-identification – especially in legal or medical contexts where faithfulness to the source mate-

rial is essential. To the best of our knowledge, this paper is the first approach specifically designed to prevent phrase-based linkage attacks while preserving the document's semantic integrity.

Record Linkage Record linkage has traditionally been studied in the context of structured datasets, where matching relies on deterministic or probabilistic comparisons of attributes such as names, birthdates, or postal codes (Christen, 2012; Kum et al., 2013; Vatsalan et al., 2013; Gkoulalas-Divanis et al., 2021). Classical linkage frameworks, such as the Fellegi–Sunter model (Fellegi and Sunter, 1969), assume tabular schemas and relatively small sets of attributes, making them ill-suited for unstructured text. Nevertheless, one key idea behind those frameworks – namely that combinations of quasi-identifiers can be as revealing as unique identifiers – carries over to text, in which infrequent phrases may act as implicit quasi-identifiers. While a few studies have examined record linkage in unstructured data, notably in the medical domain (Churi and Pawar, 2019), the problem remains underexplored. The legal implications of linkage risks in unstructured data are analysed in Weitzenboeck et al. (2022), who highlight the difficulty of complying with the GDPR's requirement of unlinkability for text or images.

Paraphrasing Paraphrasing systems have a long history in NLP, with recent progress driven by neural approaches and LLMs. These systems are typically fully automated and unsupervised (Mallinson et al., 2017; Li et al., 2018; Witteveen and Andrews, 2019; Niu et al., 2021; Wada et al., 2023). A common limitation, however, is the lack of control over which specific phrases are modified and which remain unaltered. As such, standard paraphrasing models are poorly equipped to mitigate search-based linkages. Work on controllable paraphrasing (Kumar et al., 2020; Wada et al., 2023) introduces some steerability, but are not designed to explicitly replace specific substrings for privacy protection, as is the case in the paper.

3 Approach

We first outline the threat scenario, followed by the algorithm used to identify the N-grams to rephrase and the LLM-driven rewriting procedure.

3.1 Threat model

We wish to counter privacy threats from adversaries with knowledge of at least some segments of the original documents. However, we do not know the number and types of substrings that may be available to such adversaries. To account for worst-case scenarios, we therefore adopt a maximum-knowledge intruder model (Domingo-Ferrer et al., 2015) in which the intruder has access to the original collection of documents $\{d \in \mathcal{D}\}$. Formally, given a de-identified text d', we assume the intruder seeks to establish a linkage between d' and the original text d in three consecutive steps:

- Extract one or more substrings $S_{d'} = \{s \mid s \text{ is a contiguous sequence of tokens in } d'\}$
- Search for the occurrence of those strings in $\mathcal{D}$, which results in the subset $\mathcal{D}(S_{d'})) = \{d \in \mathcal{D} \mid \forall s \in S_{d'},\ s \subseteq d\}$ of documents containing those strings.
- Finally, we consider the linkage as successful if $|\mathcal{D}(S_{d'})| < k$, where k is a minimum threshold, as done in k-anonymity (Samarati and Sweeney, 1998). We use both $k = 2$ and $k = 5$ in our experiments.

Such search-based linkages can be carried out by adversaries with minimal technical skills. As discussed in Section 5, more sophisticated types of linkages are possible, but mitigating them without incurring large distortions to the text content – and thus undermining the very aim of de-identification – is unfeasible in the general case.

3.2 Extraction of N-grams to rephrase

To efficiently determine the frequency of various N-grams (and combinations of N-grams) in a document collection $\mathcal{D}$, we rely on an *inverted index*, a data structure that maps each word or N-gram to a list (called a *posting list*) of documents it appears in. As posting lists are sorted by their document identifier, they can be efficiently intersected to derive the list of documents where a given combination of N-grams occurs (Manning et al., 2008).

Once constructed, this index can be used to determine which N-grams occurring in a de-identified document d' are associated with fewer than k documents in the original collection $\mathcal{D}$, and thus in need of rephrasing to prevent search-based linkages.

3.2.1 Single N-grams

To find a list of single N-grams to rephrase, the following procedure is applied:

1. We extract a set $S_{d'}$ of N-grams occurring in d'. To reduce the number of N-grams to consider, we limit the maximum N-gram length

(fixed to 8 words in our experiments[2]), and do not consider N-grams crossing punctuation markers or [REDACTED] PIIs, which are per definition absent from the original text d.

2. We then sort the N-grams in $S_{d'}$ by their length and filter out overlapping N-grams, resulting in a set $S_{d'}^{\min} \subseteq S_{d'}$ representing the minimal set of N-grams that need to be considered as potential linkage points[3].

3. For each N-gram $s \in S_{d'}^{\min}$, we finally check the number of documents associated with it in the inverted index $\mathcal{I}$. If the length of the posting list $\mathcal{I}(s)$ is $< k$, we add s to the N-grams to rephrase.

The result of this process is a set $S_{d'}^{*} \subseteq S_{d'}^{\min}$ of N-grams from d' that are known to be occurring $< k$ times in the original corpus of documents and which therefore need to be rephrased.

3.2.2 N-gram combinations

To extend the procedure to N-gram *combinations* instead of single N-grams, we replace step 3. above with the following two steps:

3b. We generate the set of possible N-gram combinations $C^{m}_{S_{d'}^{\min}}$ from the set $S_{d'}^{\min}$, where m is the arity of each combination. The maximum arity of the N-gram combinations was set to 3 in our experiments.

4b. For each combination $c \in C^{m}_{S_{d'}^{\min}}$, we extract the posting lists of its constituting N-grams and intersect them. As the posting lists are already sorted, this intersection can be computed in linear time. If the result of this intersection has a length $< k$, we add the shortest N-gram in c to the N-grams to rephrase.

3.3 In-context Rewriting

Once the N-grams $S_{d'}^{*}$ to rephrase in a document d' are extracted, we query an instruction-tuned LLM to generate a reformulated text in which those N-grams no longer appears *verbatim*:

$$d'_{t+1} \leftarrow \mathsf{LLM}(d'_t, S_{d'}^{*}) \quad (1)$$

The LLM is provided with the local context surrounding the N-grams to rephrase, which helps maintain fluency and contextual integration. To scale the rewriting to longer documents and ensure each LLM query is limited to a few N-grams, we chunk the text in pieces of a few sentences each. The LLM is provided with few-shot examples of rewritten sentences. We also found that instructing the LLM to generate a short chain-of-thought reasoning ahead of the rewriting led to better results. If rephrasing proves difficult – for instance, if the N-gram contains a named entity –, the LLM is instructed to replace it with a placeholder. The full prompt is provided in Appendix A.1.

As mentioned in Section 3.2, the extraction of N-grams is constrained to non-overlapping N-grams to limit the number of N-grams to consider in a single rewriting iteration. As a consequence, the rewritten text d'_{t+1} might still contain some N-grams to rephrase, as those might have been ignored due to overlaps. To this end, and to account for the fact that the LLM does not always strictly follow its instructions, the rewriting procedure is repeated for several passes until the set of N-grams $S_{d'}^{*}$ becomes empty.

4 Experiments

The approach is empirically evaluated on two datasets and along three criteria:

- *Linkage risk*: How many N-grams from the de-identified text remain linkable to the original document after rewriting?
- *Semantic integrity*: To what extent is the semantic content of the original document preserved in the rewritten text?
- *Fluency*: To what extent is the rewritten text fluent and natural?

4.1 Datasets

We rely on two datasets for evaluation, one based on court cases and the other on Wikipedia biographies. For each dataset, we start with a large collection of original documents $\mathcal{D}$, each document including various types of personally identifiable information (PII). We then draw a small subset of this collection and use a de-identification tool to produce a corresponding set of documents $\mathcal{D}'$ in which PIIs are redacted. This de-identification is practically implemented with the Presidio tool[4]. As the focus of this paper is on linkage attacks rather than PII detection, we do not evaluate here the quality of the de-identification.

[2] As the frequency of N-grams decreases sharply as a function of the N-gram length (Silva and Cunha, 2020), considering longer N-grams is usually redundant, as those would often encompass shorter N-grams that are themselves rare.

[3] We select the *shortest* non-overlapping N-grams in $S_{d'}$ to minimise the amount of editing needed to prevent linkages.

[4] See https://microsoft.github.io/presidio/

ECHR Court cases We use as first dataset $\mathcal{D}$ a set of 13 759 English-language court cases from the European Court of Human Rights (ECHR), covering the years 1961 to 2020. Court cases constitute an useful testbed for de-identification methods, due to their high density in PIIs on various individuals (offenders, victims, witnesses, etc.) and the fact that they must carefully balance privacy with the need to retain the semantic content of the case. ECHR cases have notably previously been employed in the construction of the Text Anonymisation Benchmark (TAB) (Pilán et al., 2022). A subset $\mathcal{D}'$ of 252 court cases[5] are then drawn from $\mathcal{D}$ and de-identified with Presidio.

Wikipedia biographies The second dataset consists of 100 000 biographies sampled from the English Wikipedia. The biographies were limited to medium-size lengths (between 800 and 4500 characters), and naturally contain a high density of PIIs, such as names, dates, places and organisations. As with the court cases, a subset $\mathcal{D}'$ of 250 biographies are drawn and de-identified with Presidio.

4.2 Experimental setup

As described in Section 3, the approach proceeds in several steps. An inverted index is initially constructed from the full collection of documents $\mathcal{D}$. For the ECHR dataset, the inverted index contains 58.3M distinct N-grams (up to 8 words), with an average posting list length of 1.27 documents (± 0.93). Figure 2 illustrates the distribution of those posting lists in the inverted index.

Based on this index and a de-identified document d', we derive the list of N-grams to rephrase as described in Section 3.2. Once this list is obtained, we query an instruction-tuned LLM to rewrite the text. We used in our experiments a `Qwen3-4B` model (Team, 2025), without fine-tuning but with few-shot examples to take advantage of the model's in-context learning abilities. The full prompts are provided in Appendix A.1. The responses are generated with a temperature = 1.2.

4.3 Baselines

We compare this approach to three alternative text rewriting methods:

Baseline LLM rewriting An LLM-based rewriter employing the exact same model (Qwen3-4B) and prompt as the above approach, but without being provided with a specific list of text spans to edits. In other words, the LLM is simply instructed to rewrite the full text while preserving its meaning. The rewriting process is repeated over multiple passes to ensure a sufficient number of edits compared to the original text. This baseline is used to assess the extent to which standard LLM-based paraphrasing can prevent search-based linkages.

Handcrafted paraphrasing A rule-based paraphrasing approach that replaces nouns and verbs by synonyms provided by Wordnet (Miller, 1995), along with rules for altering modifiers such as adjectives and adverbs (see Appendix A.2 for details).

DP-based rewriting The text rewriting model (Meisenbacher et al., 2024), which relies on

[5]This subset corresponds to the development and test set of the TAB corpus.

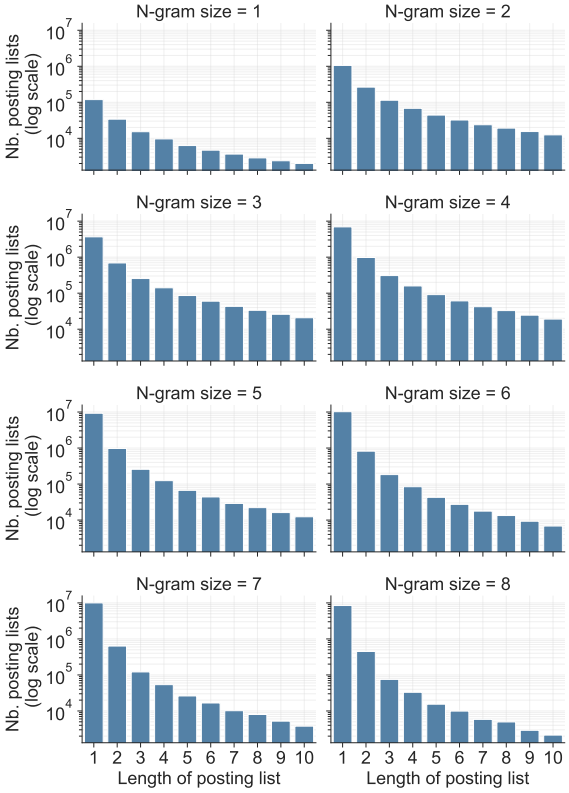

Figure 2: Distinct posting lists in the index built from the ECHR court cases, distributed according to the posting list length (=nb of documents associated with a particular N-gram) and the size of the N-gram, from 1 to 8 words. The Y-axis is displayed in log scale. Longer N-grams dominate the posting lists, as expected since the number of distinct N-grams grows combinatorially with N. The posting lists are in this case quite short, particularly for longer N-grams: 60% of posting lists for unigrams have one single document, while this proportion reaches 93% for $N = 8$.

differential privacy (DP) and masked language models (MLMs) to obfuscate the text while seeking to preserve textual similarity. We use the implementation from the authors, along with two ϵ values 10 and 100, which express the model's *privacy budget*.

4.4 Metrics

The three criteria mentioned above (linkage risk, semantic integrity and perplexity) are measured through the following metrics:

Proportion of linkable N-grams The likelihood of search-based linkages is computed by counting the number of N-grams from the document d' that can univocally link back to d, and measuring the proportion of those N-grams that remain present after rewriting. We assess both linkages with single N-grams (arity=1) as well as linkages from N-gram combinations (with an arity ≤ 3).

Semantic similarity We measure the extent to which the semantic content of d' is preserved in the rewritten text through cosine similarities between their document embeddingss, as extracted through two encoder-only language models: paraphrase-distilroberta-base-v2 and GTE-ModernBERT .

Perplexity Finally, perplexity is used to measure the linguistic quality of the rewritten document. We use the gemma 3-1B-pt base model (Gemma Team et al., 2025) for this purpose.

4.5 Results

Tables 1 and 2 present the results for the two datasets. As we can observe, the baseline LLM rewriting – which simply rewrites the full text without being provided with specific text spans to edit – preserves most of the semantic content but only hinders 40-50% of direct N-gram linkages. Handcrafted paraphrasing leads to modest numbers of edits, but with only 65-75% of direct linkages removed, and an increase in perplexity due to sub-optimal lexical choices. Finally, DP-based rewriting perform poorly, producing disfluent texts with weak semantic ties to the source. In contrast, the proposed method (last two lines in the Tables) eliminates around 99.8% of N-gram linkages, although a small number persists when the LLM is repeatedly unable to find suitable rephrasing.

We can observe a lower semantic similarity and higher perplexity of the rewritten biographies compared to the court cases. This seems to be due to the fact that many of the spans to replace in the biographies correspond to named entities or other highly specific terms that have a high impact on the document semantic similarity.

The two Tables focus on linkages with $k = 2$, which corresponds to the risk of singling out. The approach can be applied to higher values of k, at the cost of more extensive edits, leading itself to slight drops in semantic similarity[6].

Appendix A.3 provides a full, step-by-step example of a court case document rewritten using the baselines and proposed approach.

In terms of computing time, the method took (on a single GPU) an average of 15.5 ($\pm$12.8) secs per document to avert linkages with single N-grams, and 52.6 ($\pm$37.2) secs to avert linkages with N-gram combinations (up to arity 3). The court cases have an average of 750 ($\pm$389) words.

5 Discussion

Overall, the proposed approach seems to provide the most effective balance between precluding search-based linkages and preserving the semantic content of the document. The preserved content is, however, influenced by the nature of the texts: while the edits on the court cases were mostly focused on general phrases that could be easily reformulated, many edits on the Wikipedia biographies revolved around named entities and other specific, harder-to-rephrase terms, thereby leading to a larger drop in semantic similarity.

While the approach limited to the rephrasing of single "linkable" N-grams (arity=1) produces a rewritten text that is semantically closer to its input, it fails against linkages due to N-gram combinations. As shown in the results, those indirect linkages can also be removed (arity≤ 3), although at the cost of more extensive edits, leading to a slight drop in semantic similarity.

One important limitation of the presented approach is its focus on linkage attacks implemented through *phrase search*. This focus on phrase search linkages is motivated by two considerations:

- Search-based linkage attacks can be carried out by adversaries with minimal technical

[6]For instance, with $k = 5$ instead of $k = 2$, the semantic similarity of the presented approach (as measured by distilroberta on the ECHR court cases) drops from 0.987 to 0.975.

Approaches:	Proportion of linkable N-grams arity$= 1$	arity ≤ 3	Semantic Similarity distilroberta	GTE	Perplexity (Gemma3)
Baseline LLM-based rewriting (without specified spans to rephrase)	0.597	0.695	0.974	0.979	12.6
Handcrafted paraphrasing	0.348	0.615	0.992	0.979	20.4
DP-based rewriting ($\epsilon = 10$)	0.001	0.001	0.358	0.617	2736
DP-based rewriting ($\epsilon = 100$)	0.065	0.130	0.769	0.873	171
LLM-based rewriting w/ list of spans (text spans extracted for arity = 1)	0.001	0.432	0.987	0.978	14.2
LLM-based rewriting w/ list of spans (text spans extracted for arity ≤ 3)	0.002	0.002	0.959	0.948	15.0

Table 1: Evaluation results on the ECHR court cases according to three criteria: *linkage risk* (either for single N-grams or N-grams combinations of arity ≤ 3), *semantic integrity* and *fluency*. The first two columns indicate the proportion of N-grams that remain linkable to the original document d after rewriting for a value $k = 2$, which corresponds to the risk of singling out. Semantic integrity is measured through cosine similarities between the document embeddings of the text before and after rewriting. The embeddings are respectively obtained with paraphrase-distilroberta-base-v2 and GTE-ModernBERT . Finally, perplexity is measured with gemma 3-1B-pt . All LLM-based rewriting approaches are based on Qwen3-4B (see prompts in Appendix). Results obtained with a single run.

Approaches:	Proportion of linkable N-grams arity$= 1$	arity ≤ 3	Semantic Similarity distilroberta	GTE	Perplexity (Gemma3)
Baseline LLM-based rewriting (without specified spans to rephrase)	0.485	0.553	0.989	0.994	10.7
Handcrafted paraphrasing	0.256	0.355	0.952	0.978	19.17
DP-based rewriting ($\epsilon = 10$)	0.008	0.009	0.268	0.608	2122
DP-based rewriting ($\epsilon = 100$)	0.051	0.091	0.618	0.808	136
LLM-based rewriting w/ list of spans (text spans extracted for arity = 1)	0.010	0.346	0.841	0.914	12.9
LLM-based rewriting w/ list of spans (text spans extracted for arity ≤ 3)	0.002	0.003	0.733	0.858	14.7

Table 2: Evaluation results on the dataset of Wikipedia biographies according to same metrics as Table 1.

skills along with the knowledge of some text segments (N-grams) of the original texts.

- Those linkages can also be averted while retaining the core semantic content of the text.

However, more sophisticated attacks are possible. As evidenced in the results, LLM-based rewriting produce texts that remain semantically close to the initial document. While this characteristic is desirable from a data utility perspective, it also suggest that linkage attacks can be conducted at the level of *latent representations* (such as embedding vectors) instead of surface forms.

Indeed, we observed in our two datasets that when computing cosine similarities between d' and the original document collection $\mathcal{D}$, the original document d ranked first in the vast majority of cases, and that this proximity remained fairly stable across rewriting iterations, even after many rewriting passes. The proposed method does not, as a consequence, provide a robust protection against such semantics-based linkages.

Could semantics-based linkage attacks be prevented through further edits? In theory, yes, but mitigating those attacks would, by definition, re-

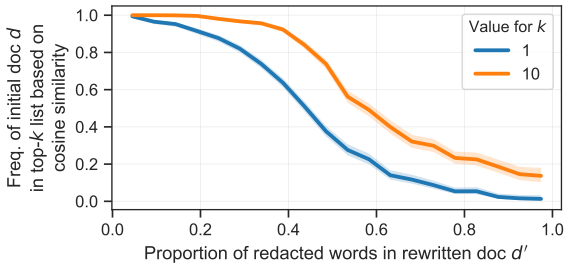

Figure 3: Frequency of original text d appearing in the top-k list among all documents in collection $\mathcal{D}$ when sorted on cosine similarity with rewritten doc d', as a function of the proportion of words redacted in d'. The results are averaged on the dataset of court cases. The shaded areas represent the Wilson CI. We can observe that the original text d remains semantically very close to the rewritten document d' until a large proportion of words (around 50%) are redacted.

quire altering the semantic content of the document to such an extent that it could no longer be related to the original text. Such large distortions would undermine the very aim of text de-identification methods, and are therefore difficult to justify from a data utility perspective.

This difficulty can be verified empirically through a simple experiment where we take a text d and gradually redact a small proportion of its words, step by step, until all words are edited out. At each iteration, we measure the cosine similarity between the initial collection $\mathcal{D}$ and the redacted text d'. We can then measure the proportion of words that need to be redacted in order to prevent semantics-based linkages – that is, to ensure the cosine similarity between d and d' is no longer among the top-k highest similarities for $\mathcal{D}$. The results for the ECHR court cases, shown in Figure 3, demonstrate that the prevention of semantics-based linkages requires drastic alterations to the document, with at least half the words needing to be redacted[7]. Needless to say, such level of obfuscation is hardly compatible with most de-identification use cases, as the resulting texts are barely readable.

However, semantics-based linkages are arguably a less pressing concern than search-based attacks. Although the threat model of Section 3.1 assumed a worst-case scenario with a maximum-knowledge intruder, search-based linkages can in practice be conducted based on the knowledge of a limited set of N-grams from the original texts. Semantics-based linkages, on the other hand, is conditioned on adversaries with access to the original documents in full text in order to compute the corresponding embedding vectors – an assumption that is unlikely to hold in practical adversarial scenarios.

[7]Alternatively, one could also increase the distance between d and d' by adding, removing or substituting segments of the text. This strategy would, however, conflict with the need to remain faithful to the original content of the text, which lies at the core of de-identification methods.

6 Conclusion

Text de-identification methods primarily focus on the detection and redaction of personally identifiable information (PII), with the objective of reducing the risk of re-identifying the individuals mentioned or referred to in the text. However, data protection regulations such as the GDPR also emphasise the need to mitigate the risk of *record linkages*, which can occur when an adversary is able to link a de-identified document to its original counterpart in the source database.

A particularly simple method of record linkage is through *phrase search*, where an adversary links the de-identified document to its original counterpart based on the occurrence of specific phrases (i.e., N-grams) that are present in the original text and remain unchanged in the de-identified version. We proposed a method to prevent such search-based linkages based on the construction of an inverted index, making it possible to determine the documents in which any particular N-gram appears. Given de-identified document d', the method extracts the N-grams appearing in fewer than k documents in the original collection and prompts an LLM to rephrase the text such that those N-grams no longer appear. The process is repeated until all problematic N-grams are rewritten.

Beyond preventing phrase linkages, the rewriting operations also acts as a broader *privacy-enhancing measure*, as they effectively limit the occurrence of rare or atypical terms in de-identified documents. Due to their rarity (relative to their underlying document collection), such distinctive terms may indeed be exploited by adversaries to infer additional cues about the potential identity of the individual(s), even when those terms do not correspond to PIIs in the narrow sense.

Experiments on two datasets demonstrate that the method can successfully mitigate the risk of search-based linkages while maintaining the core content and fluency of the rewritten texts. However, as mentioned above, more sophisticated attacks remain possible, notably based on latent semantic representations derived from the texts.

Limitations

As covered in the discussion, the proposed text rewriting approach is focused on preventing search-based linkages rather than linkages based at the level of latent representations such as document embeddings. As a consequence, the method does not prevent the possibility of semantic linkages.

Another assumption that underlies the proposed approach is that we have access to a collection $\mathcal{D}$ of original documents that constitute a "pool" from which the de-identified documents are drawn and subsequently edited. However, it should be noted that, if no such collection is available, the risk of linkages disappears, and hence the need to protect de-identified documents against those.

Due to space constraints, the evaluation of the approach is limited to two datasets, written in a single language (English). While we expect that those results also carry over other types of documents and languages beyond those datasets, this remains to be tested experimentally. The same comment applies to the use of other model architectures and parameter size than the one employed for the evaluation (Qwen3-4B).

Ethical considerations

While the proposed method aims to strengthen the privacy of de-identified texts by reducing their traceability to the original database, there is an inherent risk that such approach might give a false sense of security of the level of privacy it provides. Indeed, although de-identification can help conceal a large number of PIIs expressed in a document, they remain imperfect and prone to various errors and omissions (Sarkar et al., 2024), even when combined with the protection against search-based linkage attacks presented in this paper. It should therefore be stressed that de-identification is a privacy-enhancing technique but does not provide a strong privacy guarantee or equate full anonymisation.

References

Article 29 Data Protection Working Party. 2014. Opinion 05/2014 on anonymisation techniques. Working Party Opinion WP216, European Commission. Adopted on 10 April 2014.

Calvin Bao and Marine Carpuat. 2024. Keep it Private: Unsupervised Privatization of Online Text. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8678–8693, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter Christen. 2012. *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*, 2012 edition. Data-Centric Systems and Applications. Springer, Berlin, Germany.

Prathamesh P Churi and Ambika V Pawar. 2019. A systematic review on privacy preserving data publishing techniques. *Journal of Engineering Science & Technology Review*, 12(6).

Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *The Journal of Machine Learning Research*, 13(1):2063–2067.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association: JAMIA*, 24(3):596–606.

Tobias Deußer, Lorenz Sparrenberg, Armin Berger, Max Hahnbück, Christian Bauckhage, and Rafet Sifa. 2025. A survey on current trends and recent advances in text anonymization. In *2025 IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–9. IEEE.

Josep Domingo-Ferrer, Sara Ricci, and Jordi Soria-Comas. 2015. Disclosure risk assessment via record linkage by a maximum-knowledge attacker. In *2015 13th Annual Conference on Privacy, Security and Trust (PST)*, pages 28–35. IEEE.

Elisabeth Eder, Institut für Germanistik, Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria, Ulrike Krieg-Holz, Udo Hahn, Institut für Germanistik, Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria, and Jena University Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Jena, Germany. 2019. De-Identification of Emails: Pseudonymizing Privacy-Sensitive Data in a German Email Corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 259–269, Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.

Ivan P Fellegi and Alan B Sunter. 1969. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

Ahmed Frikha, Nassim Walha, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2025. IncogniText: Privacy-enhancing Conditional Text Anonymization via LLM-based Private Attribute Randomization. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2490–2501.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *arXiv [cs.CL]*.

Aris Gkoulalas-Divanis, Dinusha Vatsalan, Dimitrios Karapiperis, and Murat Kantarcioglu. 2021. Modern privacy-preserving record linkage techniques: An overview. *IEEE transactions on information forensics and security*, 16:4966–4987.

Sungeun Hahm, Heejin Kim, Gyuseong Lee, Hyunji M Park, and Jaejin Lee. 2025. Thunder-DeID: Accurate and efficient DE-identification framework for Korean court judgments. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12728–12755, Stroudsburg, PA, USA. Association for Computational Linguistics.

Timour Igamberdiev and Ivan Habernal. 2023. DP-BART for Privatized Text Rewriting under Local Differential Privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, page (to appear), Toronto, Canada. Association for Computational Linguistics.

Bennett Kleinberg, Toby Davies, and Maximilian Mozes. 2022. Textwash – automated open-source text anonymisation. *arXiv [cs.CL]*.

Hye-Chung Kum, Ashok Krishnamurthy, Ashwin Machanavajjhala, Michael K Reiter, and Stanley Ahalt. 2013. Privacy preserving interactive record linkage (ppirl). *Journal of the American Medical Informatics Association*, 21(2):212–220.

Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar. 2020. Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics*, 8:330–345.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2018. Paraphrase generation with deep reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3865–3878, Brussels, Belgium. Association for Computational Linguistics.

Pierre Lison, Ildikó Pilán, David Sánchez, Montserrat Batet, and Lilja Øvrelid. 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, Online. Association for Computational Linguistics.

Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, and Others. 2023. DeID-GPT: Zero-shot medical text de-identification by GPT-4. *arXiv preprint arXiv:2303. 11032*.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schutze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England.

Stephen Meisenbacher, Maulik Chevli, Juraj Vladika, and Florian Matthes. 2024. DP-MLM: Differentially private text rewriting using masked language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9314–9328, Stroudsburg, PA, USA. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

John Morris, Justin Chiu, Ramin Zabih, and Alexander Rush. 2022. Unsupervised Text Deidentification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4777–4788, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ishna Neamatullah, Margaret M Douglass, H Lehman Li-wei, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, 8(1):32.

Tong Niu, Semih Yavuz, Yingbo Zhou, Nitish Shirish Keskar, Huan Wang, and Caiming Xiong. 2021. Unsupervised paraphrasing with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5136–5150, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

David Pissarra, Isabel Curioso, João Alveira, Duarte Pereira, Bruno Ribeiro, Tomás Souper, Vasco Gomes, André Carreiro, and Vitor Rolla. 2024. Unlocking the Potential of Large Language Models for Clinical Text Anonymization: A Comparative Study. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 74–84.

P Samarati and L Sweeney. 1998. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*.

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

Atiquer Rahman Sarkar, Yao-Shun Chuang, Noman Mohammed, and Xiaoqian Jiang. 2024. De-identification is not enough: a comparison between de-identified and synthetic clinical notes. *Scientific reports*, 14(1):29669.

Joaquim F Silva and Jose C Cunha. 2020. An empirical model for n-gram frequency distribution in large corpora. In *Advances in Knowledge Discovery and Data Mining*, Lecture notes in computer science, pages 840–851. Springer International Publishing, Cham.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2024. Language Models are Advanced Anonymizers. In *The Thirteenth International Conference on Learning Representations*.

Milan Straka. 2018. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of the*, pages 197–207, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maria Irena Szawerna, Simon Dobnik, Therese Lindström Tiedemann, Ricardo Muñoz Sánchez, Xuan-Son Vu, and Elena Volodina. 2024. Pseudonymization Categories across Domain Boundaries. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13303–13314.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Dinusha Vatsalan, Peter Christen, and Vassilios S Verykios. 2013. A taxonomy of privacy-preserving record linkage techniques. *Information systems*, 38(6):946–969.

Takashi Wada, Yuji Matsumoto, Timothy Baldwin, and Jey Han Lau. 2023. Unsupervised paraphrasing of multiword expressions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4732–4746, Toronto, Canada. Association for Computational Linguistics.

Benjamin Weggenmann, Valentin Rublack, Michael Andrejczuk, Justus Mattern, and Florian Kerschbaum. 2022. DP-VAE: Human-Readable Text Anonymization for Online Reviews with Differentially Private Variational Autoencoders. In *Proceedings of the ACM Web Conference 2022*, New York, NY, USA. ACM.

Emily M Weitzenboeck, Pierre Lison, Malgorzata Cyndecka, and Malcolm Langford. 2022. The GDPR and unstructured data: is anonymization possible? *International Data Privacy Law*, 12(3):184–206.

Sam Witteveen and Martin Andrews. 2019. Paraphrasing with large language models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 215–220, Hong Kong. Association for Computational Linguistics.

Qiongkai Xu, Lizhen Qu, Chenchen Xu, and Ran Cui. 2019. Privacy-Aware Text Rewriting. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 247–257, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

A.1 Prompt

The LLM prompt employed to rephrase N-grams in a given input text (see Eq. 1) is provided below.

LLM rewriting prompt

As a text rewriting tool, you are given a text along with a list of spans occurring in that text. You must edit the text such that EVERY span in the list is replaced. The replacements must ensure that the original span no longer appears as such in the text. But the changes can be small, such as removing or inserting an adjective or adverb, using a close synonym, paraphrasing part of the span, or changing verb tense. The overall meaning should be retained as much as possible. You can also rewrite the words around the span if needed.

Many named entities have already been replaced by `[REDACTED]`, but not all. If one of the spans to replace is a named entity (person name, place, organisation, acronym, etc.) or another highly specific term impossible to rephrase, you should replace it by `[REDACTED]`. Never attempt to replace an entity by inventing a new name.

Only use `[REDACTED]` if there are really no other options to edit the span (if the span is larger than the entity, you should first edit the other words), and avoid redacting article numbers.

Examples:
Text: The quick brown fox jumps over the lazy dog.
Span(s) to replace: "quick brown fox", "over the lazy dog"
Output: The fast brown fox jumps above the lazy dog.

Text: The Kalininskiy District Court of Appeal upheld the decision of the lower court.
Span(s) to replace: "Kalininskiy District Court of Appeal", "of the lower court"
Output: The `[REDACTED]` Court of Appeal upheld the decision made by the lower court.

Text: The applicant was charged with theft and sentenced to five years in prison.
Span(s) to replace: "theft", "five years in prison"
Output: The applicant was accused of stealing and given a five-year prison term.

Text: The appeal was filed on June 5, 2020 to the District Court of Novosibirsk"
Span(s) to replace: "The appeal", "June 5, 2020", "of Novosibirsk"
Output: This appeal was filed on June 5th, 2020 to the District Court of `[REDACTED]`.

Text: The judge reviewed the evidence and ruled in favour of the prosecution.
Span(s) to replace: "reviewed the evidence", "ruled in favour of the"
Output: The judge reviewed the available evidence and decided to side with the prosecution.

Text: The judge examined the witness statements and delivered his final judgment.
Span(s) to replace: "examined the witness statements", "delivered his final"
Output: The witness statements were examined by the judge, who delivered afterwards his final judgment.

You should first output a short reasoning (max 100 words) and then output a JSON object with one single field, 'edited_text', containing the entire modified text with ALL the given spans replaced by an alternative.

Now, here is the text to edit and the spans to replace:
Text: TEXT
Span(s) to replace: SPANS_LIST

A.2 Controlled paraphrasing

One of the 3 baselines employed in the evaluation (Section 4) relies on a controlled paraphrasing approach. We provide below further technical details on this baseline.

Edit Functions The documents are initially processed using the UDPipe parser (Straka, 2018) to extract part-of-speech tags and morphological features. Edits are restricted to content words—verbs, adverbs, nouns, and adjectives. For each category, we consult the WordNet lexical database (Miller, 1995) to identify synonyms. This constitutes the first edit function, `replacement`. The second, `removal`, applies only to adjectives and adverbs. These functions are applied to each unique phrase in the development documents of the TAB corpus, generating a set of candidate replacements for any phrase requiring alteration. For `replacement`, the original word's morphological state guides the transformation of the WordNet lemma into the correct form. For instance, given the phrase *he was charged*, the lemma *charge* is searched as a verb, yielding a synonym such as *accuse*, which is then conjugated to *accused*. This process ensures correct verb forms, noun number, and adjective/adverb comparison degrees, using the Pattern Python library (De Smedt and Daelemans, 2012).

Replacement Selection To preserve semantic integrity while altering document form, we apply a weighted scoring method. Levenshtein distance quantifies form differences between original and candidate phrases. Semantic similarity is assessed using cosine similarity between embeddings generated by a sentence transformer model[8] (Zhang et al., 2019). The harmonic mean of these scores determines the optimal replacement.

A.3 Full example

A.3.1 Inputs

We provide here a full example, starting with an initial ECHR court case[9] from the document collection:

Original court case d

PROCEDURE

1. The case originated in an application (no. 5136/11) against the Republic of Poland lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms ("the Convention") by a Polish national, Mr Kamil Prus ("the applicant"), on 6 December 2010.

2. The applicant, who had been granted legal aid, was represented by Mr S. Kotuła, a lawyer practising in Lublin. The Polish Government ("the Government") were represented by their Agent, Ms J. Chrzanowska, of the Ministry of Foreign Affairs.

3. The applicant alleged, in particular, a breach of Article 3 of the Convention on account of the imposition on him of the so-called "dangerous detainee" regime.

4. On 7 July 2014 the application was communicated to the Government.

THE FACTS

I. THE CIRCUMSTANCES OF THE CASE

5. The applicant was born in 1987 and is detained in Lublin.

A. Criminal proceedings against the applicant

6. On 1 December 2005 the applicant was arrested and remanded in custody. In 2006 he was convicted of three counts of battery and robbery and received prison sentences ranging from two to four years. On 28 November 2006 the Lublin Regional Court imposed a cumulative sentence for four criminal convictions.

[8] https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens.
[9] Source: Prus v. Poland, no. 5136/11, ECHR 2016.

B. Imposition of the so-called "dangerous detainee" regime

7. On 5 November 2010 the Opole Lubelskie Prison Penitentiary Commission imposed on the applicant the so-called "dangerous detainee" regime. The commission held that the applicant had been the leader of or an active participant in a planned collective remonstrance in Opole Lubelskie Prison (Article 88a § 2 (2a) of the Code of Execution of Criminal Sentences), as the detainees refused to accept food provided by the prison authorities. The authorities learned that the detainees were planning another collective protest for 11 November 2010. The commission also considered that the applicant was particularly depraved.

8. On 18 January 2011 the Opole Lubelskie Prison Penitentiary Commission rejected an appeal lodged by the applicant and dismissed his request for leave to examine the appeal outside the statutory time-limit.

9. Subsequently, the applicant was transferred to the Lublin Remand Centre.

10. On 3 February 2011 the Lublin Remand Centre Penitentiary Commission reviewed and upheld the decision to apply the regime to the applicant, considering that he posed a serious threat to the security of the detention centre. The applicant appealed.

11. On 4 April 2011 the Lublin Regional Court dismissed the appeal. The court referred to the reasons given in the decision of 5 November 2010, holding that the decision had been lawful and justified.

12. On 3 March 2011 the applicant lodged a complaint with the Lublin Regional Inspectorate of Prisons. He complained that he was having difficulties in accessing educational activities in the Lublin Remand Centre. The authorities dismissed the complaint, holding that educational activities were not organised for dangerous prisoners.

13. On an unspecified date in 2011 the applicant notified the police of his allegedly unlawful classification as a "dangerous detainee" on 5 November 2010. He claimed that the Opole Lubelskie Prison authorities had exceeded their powers in imposing the regime on him. On 29 April 2011 the Opole Lubelskie District Prosecutor refused to open an investigation in the case, holding that no offence had been committed.

14. On 28 April 2011 the Lublin Remand Centre Penitentiary Commission revoked its decision to classify the applicant as a "dangerous detainee". It considered that the applicant's behaviour had improved and that he no longer posed a threat to the security of the remand centre.

Assume this initial court case has been de-identified, resulting in a new document d'. The de-identification was here conducted with the Presidio de-identification tool[10].

De-identified court case d'

PROCEDURE

1. The case originated in an application (no. [REDACTED]) against [REDACTED] lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms ("the Convention") by a [REDACTED] national, Mr [REDACTED] ("the applicant"), on [REDACTED] .

2. The applicant, who had been granted legal aid, was represented by [REDACTED] , a lawyer practising in Lublin. The Polish Government ("the Government") were represented by their Agent, [REDACTED] , of the Ministry of Foreign Affairs.

3. The applicant alleged, in particular, a breach of Article 3 of the Convention on account of the imposition on him of the so-called "dangerous detainee" regime.

4. On [REDACTED] the application was communicated to the Government.

[10] As one can observe in the document below, some text spans that were masked by the tool could probably have been kept in clear text (such as 'Republic of Poland') while some actual PIIs were left out of the de-identification. Those errors and omissions do not, however, affect in any way the operation of the proposed approach.

THE FACTS

I. THE CIRCUMSTANCES OF THE CASE

5. The applicant was born in [REDACTED] and is detained in [REDACTED].

A. Criminal proceedings against the applicant

6. On [REDACTED] the applicant was arrested and remanded in custody. In [REDACTED] he was convicted of three counts of battery and robbery and received prison sentences ranging from [REDACTED]. On [REDACTED] the Lublin Regional Court imposed a cumulative sentence for four criminal convictions.

B. Imposition of the so-called "dangerous detainee" regime

7. On [REDACTED] the Opole Lubelskie Prison Penitentiary Commission imposed on the applicant the so-called "dangerous detainee" regime. The commission held that the applicant had been the leader of or an active participant in a planned collective remonstrance in [REDACTED] (Article 88a § 2 (2a) of the Code of Execution of Criminal Sentences), as the detainees refused to accept food provided by the prison authorities. The authorities learned that the detainees were planning another collective protest for [REDACTED]. The commission also considered that the applicant was particularly depraved.

8. On [REDACTED] the Opole Lubelskie Prison Penitentiary Commission rejected an appeal lodged by the applicant and dismissed his request for leave to examine the appeal outside the statutory time-limit.

9. Subsequently, the applicant was transferred to the Lublin Remand Centre.

10. On [REDACTED] the Lublin Remand Centre Penitentiary Commission reviewed and upheld the decision to apply the regime to the applicant, considering that he posed a serious threat to the security of the detention centre. The applicant appealed.

11. On [REDACTED] the Lublin Regional Court dismissed the appeal. The court referred to the reasons given in the decision of [REDACTED], holding that the decision had been lawful and justified.

12. On [REDACTED] the applicant lodged a complaint with the Lublin Regional Inspectorate of Prisons. He complained that he was having difficulties in accessing educational activities in the Lublin Remand Centre. The authorities dismissed the complaint, holding that educational activities were not organised for dangerous prisoners.

13. On an unspecified date in [REDACTED] the applicant notified the police of his allegedly unlawful classification as a "dangerous detainee" on [REDACTED]. He claimed that the Opole Lubelskie Prison authorities had exceeded their powers in imposing the regime on him. On [REDACTED] the Opole Lubelskie District Prosecutor refused to open an investigation in the case, holding that no offence had been committed.

14. On [REDACTED] the Lublin Remand Centre Penitentiary Commission revoked its decision to classify the applicant as a "dangerous detainee". It considered that the applicant's behaviour had improved and that he no longer posed a threat to the security of the remand centre.

A.3.2 Proposed approach

Based on the approach presented in Section 3.2, we can use the inverted index to derive the N-grams occurring only once or twice ($k = 2$) in the document collection:

N-grams to rephrase (< 2 occurrences in $\mathcal{D}$)

['battery and robbery', 'and received prison sentences ranging', 'Lublin Regional Court imposed', 'for four criminal', 'Lubelskie Prison Penitentiary Commission imposed', 'The commission held that the applicant had', 'leader of or', 'active participant in a', 'the detainees refused', 'accept food

provided', 'The authorities learned', 'planning another', 'collective protest for', 'commission also considered', 'particularly depraved', 'Commission rejected an', 'by the applicant and dismissed his', 'for leave to examine', 'the appeal outside the', 'reviewed and upheld the decision', 'apply the regime', 'that he posed a serious', 'security of the detention', 'court referred to the reasons given in the', 'Inspectorate of Prisons', 'he was having difficulties in', 'accessing educational', 'activities in the Lublin', 'The authorities dismissed', 'holding that educational', 'not organised for', 'police of his allegedly', 'unlawful classification', 'detainee" on', 'claimed that the Opole', 'Lubelskie Prison authorities', 'powers in imposing', 'Lubelskie District Prosecutor refused', 'holding that no offence', 'Commission revoked its', 'It considered that the applicant's behaviour', 'improved and that he no']

After querying the LLM to rewrite the N-grams above, we obtain the following results:

De-identified court case d' after rewriting N-grams with < 2 occurrences in $\mathcal{D}$

PROCEDURE

1. The case originated in an application (no. [REDACTED]) against [REDACTED] lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms ("the Convention") by a [REDACTED] national, Mr [REDACTED] ("the applicant"), on [REDACTED].

2. The applicant, who had been granted legal aid, was represented by [REDACTED], a lawyer practising in Lublin. The Polish Government ("the Government") were represented by their Agent, [REDACTED], of the Ministry of Foreign Affairs.

3. The applicant alleged, in particular, a breach of Article 3 of the Convention on account of the imposition on him of the so-called "dangerous detainee" regime.

4. On [REDACTED] the application was communicated to the Government.

THE FACTS

I. THE CIRCUMSTANCES OF THE CASE

5. The applicant was born in [REDACTED] and is detained in [REDACTED].

A. Criminal proceedings against the applicant

6. On [REDACTED] the applicant was arrested and remanded in custody. In [REDACTED] he was convicted of three counts of assault and theft and was handed a prison sentence ranging from [REDACTED]. On [REDACTED] the Lublin Regional Court handed down a cumulative sentence across four offenses.

B. Imposition of the so-called "dangerous detainee" regime

7. On [REDACTED] the Opole Lubelskie Prison Penitentiary Commission ordered on the applicant the so-called "dangerous detainee" regime. The commission concluded that the applicant had been the head of or involved in a coordinated group action in [REDACTED] (Article 88a § 2 (2a) of the Code of Execution of Criminal Sentences), as the detainees declined the meals supplied by the prison authorities. The authorities discovered that the detainees were organizing a new group protest for [REDACTED]. The commission also noted that the applicant showed extreme insensitivity.

8. On [REDACTED] the body declined to accept an appeal submitted by the applicant and rejected the applicant's request to review the appeal beyond the statutory deadline.

9. Subsequently, the applicant was transferred to the Lublin Remand Centre.

10. On [REDACTED] the Lublin Remand Centre Penitentiary Commission considered and confirmed the decision to impose the custody conditions for the applicant, who was deemed a major threat to the detention facility's safety. The applicant appealed.

11. On [REDACTED] the Lublin Regional Court dismissed the appeal. The court considered the

grounds presented in the decision of [REDACTED], stating that the decision had been lawful and reasonable.

12. On [REDACTED] the applicant submitted a complaint to the [REDACTED] of Prisons. He expressed that he faced challenges in engaging in educational programs at the Lublin Remand Centre. The officials rejected the complaint, asserting that educational programs were not provided for dangerous prisoners.

13. On an unspecified date in [REDACTED] the applicant notified the authorities of his allegedly illegitimate designation as a "dangerous individual" on [REDACTED]. He asserted that the [REDACTED] authorities had overstepped their responsibilities in applying the regime to him. On [REDACTED] the [REDACTED] refused to initiate an investigation, stating that no violation had occurred.

14. On [REDACTED] the [REDACTED] withdrew its determination to classify the applicant as a "dangerous individual." It noted that the applicant's conduct had changed and that he no longer presented a risk to the safety of the remand center.

Finally, one can push the rewriting even further and also query the LLM to rephrase all *combinations* of N-grams (up to a maximum arity of 3) that only occur once or twice in the collection. This rewriting further enhances the robustness of the text against search-based linkage attacks, although at the cost of more sweeping changes to the content of the document:

De-identified case d' after rewriting N-gram combinations of arity $\in \{2,3\}$

PROCEDURE

1. The case originated in an application (no. [REDACTED]) against [REDACTED] lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms("the Convention") by a [REDACTED] national, Mr [REDACTED] ("the applicant"), on [REDACTED] .

2. The applicant, who had been granted legal aid, was represented by [REDACTED], a lawyer practising in [REDACTED]. The Polish Government("the Government") were represented by their Agent, [REDACTED], of the Ministry of Foreign Affairs.

3. The applicant alleged, in particular, a breach of Article 3 of the Convention on account of the imposition on him of the so-called [REDACTED] regime.

4. On [REDACTED] the application was communicated to the Government.

THE FACTS

I. THE CIRCUMSTANCES OF THE CASE

5. The applicant was born in [REDACTED] and is detained in [REDACTED].

A. Criminal proceedings against the [REDACTED]

6. On [REDACTED] the [REDACTED] was arrested and remanded in custody. In [REDACTED] he was convicted of three counts of assault and theft and was handed a prison sentence spanning from [REDACTED]. On [REDACTED] the [REDACTED] Court handed down a total sentence across four offenses.

B. Application of the so-called "high-risk detainee" regime

[REDACTED]. On [REDACTED] the [REDACTED] Court ordered on the individual the so-called "high-risk detainee" regime. The body concluded that the applicant had been the head of or involved in a coordinated group action in [REDACTED] (Article [REDACTED] § 2 (2a) of the Code of Execution of Criminal Sentences), as the individuals declined the meals supplied by the prison authorities. The authorities discovered that the individuals were organizing a new demonstration for [REDACTED]." The body also noted that the applicant showed extreme insensitivity.

8. On [REDACTED] the body declined to accept an appeal submitted by the applicant and rejected the

applicant's request to review the appeal beyond the statutory deadline.

9. Subsequently, the applicant was transferred to the [REDACTED].

10. On [REDACTED] the [REDACTED] Commission considered and confirmed the decision to impose the custody conditions for the applicant, who was deemed a major threat to the detention facility's safety. The applicant appealed.

11. On [REDACTED] the [REDACTED] Court dismissed the appeal. The court considered the grounds presented within the ruling of [REDACTED], stating that the decision had been lawful and reasonable.

12. On [REDACTED] the applicant submitted a complaint to the [REDACTED] of [REDACTED]. He expressed that he faced challenges in engaging in education at the [REDACTED] Remand Centre. The officials rejected the complaint, asserting that education was not provided for high-risk prisoners.

13. On an unspecified date in [REDACTED] the applicant notified the authorities of his allegedly illegitimate designation as a "dangerous individual" on [REDACTED]. He asserted that the [REDACTED] authorities had overstepped their responsibilities in the application of the system against him. On [REDACTED] the [REDACTED] refused to initiate an investigation, stating that no violation had occurred.

14. On [REDACTED] the [REDACTED] withdrew its determination to designate the applicant as a "dangerous individual." It noted that the applicant's conduct had changed and that he no longer presented a risk to the safety of the remand center.

In this last example, the LLM also replaced some N-grams with [REDACTED] when those were impossible to paraphrase. This is particularly the case for named entities such as organisations.

A.3.3 LLM rewriting without specified spans to rephrase

As discussed in Section 4, a straightforward baseline involves prompting the LLM to rewrite the de-identified document without supplying the specific N-grams to be rephrased. The resulting text was generated over three rewriting iterations, using the same prompt as outlined in Section A.1, but excluding the span specifications.

De-identified court case d' after LLM rewriting without specified spans to rephrase

PROCEDURE

1. The case began with a claim (no. [REDACTED]) against [REDACTED] submitted to the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms ("the Convention") by a [REDACTED] national, Mr [REDACTED] ("the applicant"), on [REDACTED].

2. The applicant, who had been granted legal aid, was represented by [REDACTED], a lawyer based in Lublin. The Polish Government ("the Government") was represented by their lawyer, [REDACTED], of the Ministry of Foreign Affairs.

3. The applicant alleged, in particular, a breach of Article 3 of the Convention on account of the imposition on him of the so-called "dangerous detainee" regime.

4. On [REDACTED] the application was communicated to the Government.

THE FACTS

I. THE CIRCUMSTANCES OF THE CASE

5. The applicant was born in [REDACTED] and is detained in [REDACTED].

A. Criminal proceedings against the applicant

6. On [REDACTED] the applicant was arrested and remanded in custody. In [REDACTED] he was convicted of three counts of battery and robbery and received prison terms ranging from [REDACTED]. On [REDACTED] the Lublin Regional Court imposed a cumulative sentence for four criminal convictions.

B. Imposition of the so-called "dangerous detainee" regime

7. On [REDACTED] the Opole Lubelskie Prison Penitentiary Commission imposed on the applicant the so-called "dangerous detainee" regime. The commission determined that the applicant had been the leader or an active participant in a planned collective remonstrance in [REDACTED] (Article 88a § 2 (2a) of the Code of Execution of Criminal Sentences), as the detainees refused the food provided by the prison authorities. The authorities learned that the detainees were planning another collective protest for [REDACTED]. The commission also considered that the applicant was particularly egregious.

8. On [REDACTED] the [REDACTED] rejected an appeal filed by the applicant and denied his request to examine the appeal outside the statutory time-limit.

9. Subsequently, the applicant was moved to the [REDACTED].

10. On [REDACTED] the [REDACTED] examined and confirmed the decision to apply the regime to the applicant, considering that he posed a serious threat to the security of the detention centre. The applicant appealed.

11. On [REDACTED] the [REDACTED] dismissed the appeal. The court cited the grounds outlined in the ruling, holding that the decision was lawful and reasonable.

12. On [REDACTED] the applicant filed a grievance with the [REDACTED] of Prisons. He struggled to access learning opportunities at the [REDACTED] Centre. The authorities dismissed the complaint, stating that learning opportunities were not provided for dangerous prisoners.

On an unspecified date in [REDACTED] the applicant informed the police of his asserted unlawful classification as a "dangerous detainee" on [REDACTED]. He asserted that the Opole Lubelskie Prison authorities had improperly imposed the regime on him. On [REDACTED] the Opole Lubelskie District Prosecutor refused to initiate an inquiry in the case, holding that no offence had been committed.

14. On [REDACTED] the [REDACTED] revoked the decision to classify the applicant as a "dangerous detainee." It determined that the applicant's behaviour had improved and that he no longer posed a risk to the security of the remand centre.

While numerous sentences in the input document have been paraphrased, the rewriting appears largely indiscriminate – many frequent N-grams are altered unnecessarily, whereas the majority of the rare N-grams (22 out of 30) leading to linkages with the original document d remain unchanged.

A.3.4 Controlled paraphrasing

As detailed in Appendix A.2, we also implemented an paraphrasing approach relying on hancrafted rules. More specifically, the approach first parsed the document to obtain POS tags and morphological features. Nouns and verbs were then replaced by synonyms, while heuristic rules were employed to add or delete modifiers such as adjectives and adverbs.

The outcome of this approach on our example is as follows:

De-identified court case d' after using controlled paraphrasing

PROCEDURE

1. The case originated in an application (no. [REDACTED]) against [REDACTED] lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms ("the Convention") by a [REDACTED] national, Mr [REDACTED] ("the applicant"), on [REDACTED].

2. The applicant, who had been granted legal aid, was represented by [REDACTED], a lawyer practising in Lublin. The Polish Government ("the Government") were represented by their Agent, [REDACTED], of the Ministry of Foreign Affairs.

3. The applicant alleged, in particular, a breach of Article 3 of the Convention on account of the imposition on him of the so-called "life-threatening detainee" authorities.

4. On [REDACTED] the application was communicated to the Government.

THE FACTS

I. THE CIRCUMSTANCES OF THE CASE

5. The applicant was born in [REDACTED] and is detained in [REDACTED] .

A. Criminal proceedings against the applicant

6. On [REDACTED] the applicant was arrested and remanded in custody. In [REDACTED] he was convicted of three counts of bombardment and looting and received prison house sentences ranging from [REDACTED] . On [REDACTED] the lublin court imposed a cumulative time for four felonious convictions.

B. infliction of the so-called “ unsafe political detainee ” authorities

7. On [REDACTED] the Opole Lubelskie Prison Penitentiary Commission imposed on the applicant the so-called “ unsafe detainee” regime. the direction held that the applier had been the leader of or an participating player in a planned remonstration in [REDACTED] (Article 88 amp § 2 (2a) of the Code of Execution of Criminal Sentences), as the detainees refused to have food for thought provided by the prison house authorities. The authorities learned that the detainees were planning another protest for [REDACTED] . The delegation also considered that the applicant was in particular perverted.

8. On [REDACTED] the Opole Lubelskie Prison Penitentiary Commission rejected an appeal lodged by the applier and dismissed his petition for leave of absence to prove the solicitation outside the statutory time-limit.

9. Subsequently, the applicant was transferred to the Lublin Remand Centre.

10. On [REDACTED] the Lublin Remand Centre Penitentiary Commission reviewed and upheld the decisiveness to implement the authorities to the applicant, considering that he posed a sober threat to the protection of the detention centre. The applicant appealed.

11. On [REDACTED] the lublin Court dismissed the appeal. The judicature referred to the reasons given in the decisiveness of [REDACTED] , holding that the decision had been lawful and justified.

12. On [REDACTED] the applicant lodged a complaint with the lublin Inspectorate of Prisons. He complained that he was having difficulties in accessing activities in the Lublin Remand Centre. The authorities dismissed the complaint, holding that activities were not unionized for dangerous prisoners.

13. On an unspecified date in [REDACTED] the applicant notified the police of his wrongful assortment as a “ unsafe political detainee ” on [REDACTED] . He claimed that the Opole Lubelskie Prison authorities had exceeded their powers in imposing the regimen on him. On [REDACTED] the Opole Lubelskie District Prosecutor refused to open an investigation in the case, holding that no discourtesy had been committed.

14. On [REDACTED] the Lublin Remand Centre Penitentiary Commission revoked its decision to sort out the applicant as a “life-threatening detainee”. it considered that the applicant’s demeanor had improved and that he no longer posed a threat to the security of the remand centre.

We can observe that the approach is able to maintain the semantic content of the text, although a substantial number of linkable N-grams remain present. In addition, the lack of context-sensitivity of the rewriting mechanism does sometimes lead to suboptimal or wrong replacements, such as "battery and robbery" being replaced by "bombardment and looting".

A.3.5 DP-based rewriting

One final alternative is to rely on approaches grounded in Differential Privacy (DP). We have notably experimented with the DPMLM model of (Meisenbacher et al., 2024). Unfortunately, the quality of the resulting documents is very poor (here with an $\epsilon = 100$):

De-identified court case d' after DP-based rewriting with ϵ=100

State 5. The cases originated in an application (no. [Red]) against [Red] lodged with the Judge under Article 25 of the Conference for the Conservation of Human Right and Human Issues (" the Convention () by a [Acted] national, Mr [Red] (" the applicant (), on [Red].

3. This applicant, who had been received fraudulent aiding, was testified by [Red], a comedian pract in Warsaw. That Greek State (" the Governments?) were featured by their Agents, [Acted], of the Government of Foreign Arts.

4. They application did, in particulars, a break of Articles iii of the Convention on aim of the implementation on him of the the " dangerous detainee regime.

4. In [Red] the application was given to the Government. Related Facts O. The Background Are Your Court

8. The petitioner was born in [Red] and is held in [Red]. A. Criminal cases against the appellant

6. Off [Red] the respondent was handcuffed and put in custody. In [Red] he was condemned of 43 monsters of battery and robbers and earned incarcerated orders scoring from [Red]. After [Red] the Moscow R Department announced a cumulative suspension for four different punishments.

6. Introduction of the alleged " dangerous detainee dictatorship

9. On [Red] the Op K Prison Pen Commission initiated on the applicant the " dangerous detainee dictatorship. The commission withheld that the applicants had been the leader of or an advanced plaintiff in a packed collectively demonstration in [Redacted] (Articles 88 § 3 (2) of the Title of Administration of Criminal Ences), as the patients determined to permit material released by the jail complexes. Some courts suspected that the occupants were practicing a collectively confrontation for [Red]. This commissioner apparently deemed that the respondent was additionally sexist.

3. And [Red] the Ole C Province Prison Commission revoked an intervention received by the holder and released his attempt for permission to view the appeal after the contractual limitation.

14. Upon, the applicant was copied to the Munich Reform Center.

14. And [Red] the Munich Detention Centre Prison Commission appealed and approved the proposal to grant the regime to the appellant, forwarding that he issued a deep menace to the health of the incarceration room. They holder appeals.

11. On [Red] the Warsaw Superior Cour revived the argument. The court referring to the examples passed in the declaration of [Red], holding that the case had been just and decisive.

3. After [Acted] the application directed a complain with the Berlin Rural Inventory of Isons. Once said that he was having supports in attending educational facilities in the Berlin Detention Centre. State authorities closed the matter, hold that instructional events were not meant for serious ones.

28. Before an mentioned date in [Red] the applicants briefed the police of his unbelievably lawful listing as a « dangerous inmate on [Acted]. Also suggested that the Op Kie Prison apparatus had unleashed their policies in assuming the regime on him. When [Acted] the Warsaw Cki Public Prosecutor ruled to launch an exhibition in the cases, restraining that no offence had been confessed.

19. : [Red] the Berlin And Centre Correctional Committee repealed its decision to exclude the applicants as a " dangerous offender ". In reached that the attacker? s act had changed and that he no again ranked a challenge to the quality of the and centre.

975