Re-identification of De-identified Documents with Autoregressive Infilling

Anonymous ACL submission

Abstract

Documents revealing sensitive information about human individuals must often be de-identified prior to being released. This de-identification is typically done by masking all mentions of personal identifiers, thereby making it more difficult to uncover the identity of the person(s) in question. To investigate the robustness of de-identification methods, we present a novel, RAG-inspired approach that attempts the reverse process of *re-identification* based on a database of documents representing background knowledge. Given a de-identified text in which personal identifiers have been masked, the re-identification proceeds in two steps. A retriever first selects from the background knowledge passages deemed relevant for the re-identification. Those passages are then provided to an infilling model which seeks to infer the original content of each text span. This process is repeated until all masked spans are replaced. We evaluate the re-identification on two datasets based on Wikipedia biographies and court cases. Results show that (1) as many as 80% of de-identified text spans can be successfully recovered and (2) the re-identification accuracy increases along with the level of background knowledge.

1 Introduction

Many types of text documents contain sensitive information about human individuals, including e.g. clinical notes, court cases or email interactions with social services. When those documents need to be published or transferred to third parties, it is typically desirable – and sometimes legally required – to *de-identify* them beforehand. Most de-identification approaches operate by (1) determining the text spans that express direct or indirect personal identifiers and (2) masking those from the document. This process can be done manually or using NLP models (Sweeney, 1996; Neamatullah et al., 2008; Sánchez and Batet, 2016; Dernoncourt et al., 2017; Lison et al., 2021; Liu et al., 2023).

It is, however, difficult to properly assess whether the de-identification has adequately concealed the identity of the person(s) mentioned in the original document. Many evaluation techniques assess the performance of de-identification methods by comparing their outputs with those of human experts (Lison et al., 2021; Pilán et al., 2022). However, those evaluation techniques depend on the availability of human annotations and may be prone to human errors and inconsistencies.

An alternative approach to evaluating the de-identification performance is through an automated *adversary* that attempts to infer the original context of each text span that had been masked (Mozes and Kleinberg, 2021; Manzanares-Salor et al., 2022). This paper presents such an adversarial approach, based on a retrieval-augmented scheme where relevant information is first retrieved from a body of background knowledge, and then exploited to infer the original content that hides behind each masked text span. The background knowledge should ideally represent all information that one may assume will be available to adversaries. As shown by the evaluation results, the amount of information included as background knowledge notably influences the re-identification accuracy.

The rest of the paper is as follows. Section 2 introduces the relevant background on text de-identification, text infilling and retrieval-augmented methods. Section 3 describes the re-identification approach, which is then evaluated in Section 4 on two datasets. Finally, Section 5 and Section 6 discuss the results and outline future directions.

2 Background

2.1 Text de-identification

Personal data is protected through several legal frameworks, such as the European General Data Protection Regulation (GDPR, 2016) introduced in 2018. An important principle outlined in those le-

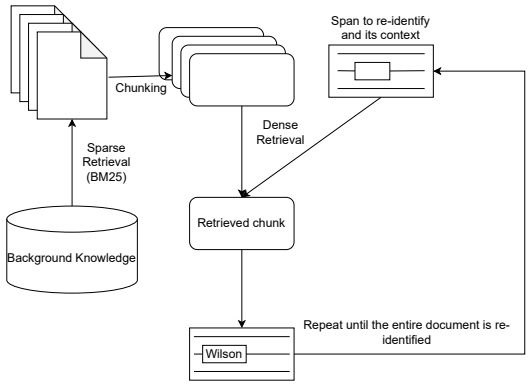

Figure 1: Sketch of the re-identification pipeline. A sparse retriever first selects the k most relevant documents from the background knowledge, and a dense retriever then extracts the most relevant chunk from those documents. Finally, the infilling model generates a possible re-identification given the context and the retrieved chunk.

gal frameworks is *data minimization*, which states that data owners should restrict the data collection and processing to only what is required to fulfill a specific purpose. The goal of text de-identification, also called text sanitization (Sánchez and Batet, 2016; Papadopoulou et al., 2022), is precisely to fulfill this data minimization principle by making it more difficult to re-identify the person from the text (Lison et al., 2021; Pilán et al., 2022).

Personally identifiable information, or PII, can be divided into two categories, both of which should be masked from the text to ensure the texts are properly de-identified (Elliot et al., 2016):

Direct identifiers, which are defined as information that can univocally identify an individual, such as the person's name, phone number, home address or passport number.

Quasi identifiers, which are not *per se* sufficient to single out an individual, but may do so when combined with one another. Examples of quasi-identifiers include the person's nationality, occupation, gender, place of work, date of birth or physical appearance.

Evaluating de-identification methods is a challenging task. A common solution is to compare the outputs against manually annotated documents (Pilán et al., 2022). Relying on manual annotations is, however, not always feasible, and is hampered by the presence of residual errors, omissions, and inconsistencies in those human judgments. One alternative is to carry out re-identification attacks on the de-identified documents to determine whether an adversary could uncover the identity of the person to protect (Scaiano et al., 2016; Mozes and Kleinberg, 2021). Notably, Manzanares-Salor et al. (2022) train a neural text classifier to link back Wikipedia biographies with its corresponding person name. This classifier, however, directly predicts the person's name from the text. In contrast, the approach present in this paper takes advantage of the generic background knowledge encoded in large language models to first uncover the masked text spans and only seeks to predict the person's identity after this unmasking step.

The idea of constructing an adversary seeking to unveil a sensitive attribute has also been explored in the area of text rewriting (Xu et al., 2019). However, those approaches typically seek to protect other attributes than the person's identity (such as gender or ethnicity) and focus on different types of document edits than the masking of PII. Such complete transformations of the text can also performed using methods based on differential privacy (Krishna et al., 2021; Igamberdiev and Habernal, 2023), although those methods do not typically conduct explicit re-identification attempts.

2.2 Text infilling

The problem of predicting missing spans of text at any position within a document (often indicated via a special placeholder symbol) is known as *infilling* (Zhu et al., 2019; Donahue et al., 2020) or *fill-in-the-middle* (Bavarian et al., 2022). In contrast to masked language models such as BERT (Devlin et al., 2019), which are pretrained to predict a single masked token based on the surrounding context, the infilling task may span multiple tokens (whose number is typically left unknown, although one can control its length). Two early approaches to text infilling were respectively presented by Zhu et al. (2019) and Donahue et al. (2020). Those two approaches demonstrated how to use pretraining and fine-tuning to enable a language model to fill in spans of a controlled size. More recently, a Generalized Language Model (GLM) was proposed by Du et al. (2022), comprising both encoder and decoder architectures. For decoder models, it combines the standard autoregressive task with the infilling task by giving the ability for the model to be bidirectional before the generation marker. For encoder models, GLM generalizes the standard token-level masking problem by (1) masking entire spans with a single token and (2) training the model to autore-

gressively generate the correct replacement span at the end of the text.

2.3 Retrieval-augmented models

The factual knowledge stored in standard language models is distributed among all model parameters and cannot be easily edited, updated, or even inspected. *Retrieval-augmented language models* (Lewis et al., 2020; Guu et al., 2020; Ram et al., 2023) seek to address this shortcoming by coupling the model with a knowledge base of documents. The generation process is then split into a *retrieval* phase, in which relevant documents from the knowledge base are extracted, and a *reading* phase, which corresponds to the actual generation, conditioned on both the context and the relevant documents selected by the retriever.

Retrieval-augmented systems make it possible to edit, extend or update the knowledge base of documents while keeping the underlying language model unchanged (Gao et al., 2023). The retrieval mechanism can also enhance the system's interpretability, as one can directly inspect the retrieved documents and assess their influence in the final output of the model (Sudhi et al., 2024).

There are multiple ways to train retrieval-augmented models. A common strategy is to rely on pre-trained retriever and reader models, and then fine-tune those two end-to-end on a standard language modelling objective, as shown in e.g. (Lewis et al., 2020). Another approach is to continue pre-training of the language model with a retriever that could be trained (Guu et al., 2020) or not (Izacard et al., 2023). Language models trained from scratch with a trained retriever have also been proposed (Borgeaud et al., 2022).

The approach described in this paper is directly inspired by Retrieval-Augmented Generation (RAG) methods, as we also seek to improve the prediction performance with the help of a neural retriever connected with a knowledge base. However, while most previous work on RAG has concentrated on tasks such as question answering, we focus here on the task of re-identifying a document in which personal identifiers have been masked.

3 Approach

The proposed method is divided into two main steps. Given a de-identified document and a particular masked span which we seek to uncover, we first *retrieve* a list of relevant passages from a database of background documents. Using those passages, a fine-tuned LLM then generates *infilling hypotheses* for the masked span. The operation is repeated until all masked spans in the de-identified document are replaced by their most likely hypothesis. We describe those steps below.

3.1 Retrieval

The retriever model relies on a database of documents representing the background knowledge available for the re-identification. This background knowledge should ideally comprise all information that one can expect to be available to an adversary seeking to uncover the personal information that the de-identification sought to conceal.

As this background knowledge will often be quite large, we decompose the retrieval process in two separate steps. A *sparse retriever* is first employed to find relevant background documents for the de-identified text. As some of those documents may be particularly long and include many irrelevant parts, we then apply a *dense retrieval* model to determine, within each document, the passages that are most relevant to unmask a particular span in the de-identified text.

Sparse document retriever

The sparse retriever takes as input a de-identified text and outputs a list of relevant documents from the background knowledge. To efficiently search for those documents, we rely on the BM25 algorithm (Robertson et al., 2009) with a default setup and retrieve the N most similar documents (where N was set to 10 in our experiments).

Dense passage retriever

The documents selected by the sparse BM25 retriever are then split into overlapping chunks of about 600 characters each. For each masked span in the de-identified document, we create a query string of 128 tokens consisting of the local context around that span. The masked span in that query is denoted with a special `[ANON]` token.

The dense retriever is a fine-tuned ColBERT model (Khattab and Zaharia, 2020). The data employed for the fine-tuning consists of both positive and negative (passage, query) pairs. The positive pairs are defined as passages that include the original content of the span that was masked, while the negative pairs are passages that do not. For instance, if the sentence "The applicant was born in the German city of Aachen" was de-identified

as "The applicant was born in the German city of [ANON]", the pair ("Aachen is the westernmost city in Germany', "The applicant was born in the German city of [ANON]") will constitute a positive example for the retriever. This setup makes it possible to fine-tune the ColBERT retriever model independently of the infilling model.

3.2 Infilling

The passages deemed as relevant by the retriever are then employed to re-identify each masked span in the de-identified document. In our experiments, the number of passages included for the infilling was set to either 1 or 2. Next to those passages, we also provide the actual context of the span we seek to re-identify, such as "The applicant was born in the German city of [ANON]".

We experiment with two distinct LLMs to generate hypotheses for this infilling task. The first is a GLM RoBERTA Large model (Du et al., 2022), where the context is provided using a 200-character window to the left and the right of the span. We also experiment with a Phi-1.5 model (Li et al., 2023) that is given a 300-character left-side context.

While we could in principle use those LLMs to generate infilling hypotheses without fine-tuning, we found that fine-tuning on in-domain data improved the infilling results, as it incites the LLM to exploit the information provided in the retrieved passages in addition to the context of the span.

Given a de-identified document, we replace each masked span one at a time, in randomized order, until all masked spans are replaced by the infilling model. Masked spans that are not yet replaced but are not the focus of the current infilling are replaced with the special [ANON] token.

4 Evaluation

The re-identification method is evaluated on two distinct datasets. The first one is a generic corpus extracted from Wikipedia in which personal identifiers have been masked using a standard Named Entity recognizer, while the second is the Text Anonymization Benchmark (Pilán et al., 2022), which was explicitly designed for privacy-oriented NLP tasks, and has been manually annotated with both direct and quasi-identifiers.

To assess more precisely the extent to which the background knowledge influences the re-identification performance, we evaluate the method with four levels of background knowledge:

Level 1: No retrieval : In this setup, no background knowledge is assumed and the infilling is directly performed by the generation model.

Level 2: General knowledge : We include texts that might be relevant for the re-identification, but without including similar texts.

Level 3: All texts except document : This setup extends the database of general knowledge with similar documents, but without including the text we seek to re-identify.

Level 4: All texts including document : This setup mimics a strong adversary who has access to background documents including the original version of the text to re-identify.

4.1 Data

Wikipedia Biographies

The Wikipedia biographies dataset consists of all biographies found on Wikipedia identified by the Biography WikiProject.[1] This represents 2 001 380 biographies. This dataset is used to train the model and create a synthetic re-identification dataset. To sanitize the biographies we use an English NER model from Spacy[2] and remove every named entity identified by the model.[3]

For our general knowledge, we use the rest of Wikipedia (i.e. non-biographies) which represents 4 732 020 articles. These articles could relate to e.g. discoveries or events connected to the person referred to in the biography. For the levels 3 and 4 above, we also include the Wikipedia biographies themselves, respectively without and with the actual biography to re-identify.

Text Anonymization Benchmark (TAB)

The TAB dataset (Pilán et al., 2022) consists of 1 268 English-language court cases from the European Court of Human Rights (ECHR). Each court case has been manually de-identified and includes detailed annotations such as identifier type, semantic category and confidential attributes.

Level 2 of background knowledge is compiled from a collection of 28 569 legal summaries, reports, and communicated cases from the ECHR.

[1] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Biography

[2] https://spacy.io/models/en#en_core_web_trf

[3] Although not all named entities are personal identifiers, and personal identifiers may also correspond to expressions that are not named entities, there is a strong correlation between the two, especially in Wikipedia biographies.

These contain similar language and case information that could help the re-identification. Levels 3 and 4 also include the court cases themselves.

4.2 Training details

Retrieval

To train the ColBERT model employed for the dense retrieval, we use the sanitized Wikipedia biographies and the non-biographies as databases. After splitting the documents into text chunks we create a dataset to train our ColBERT model with positive examples being those that contain the span to re-identify and the negatives not containing it.

We fine-tune the dense retriever from the ColBERT model for English, more precisely two case-sensitive base-sized BERTs for embedding the documents and queries. We train the model for 10 000 steps with a batch size of 128 and compress each document and query token to dimension 32 from 768. Finally, as in (Khattab and Zaharia, 2020), we fix the sequence length of the queries to 128 tokens and use the extra tokens as "memory tokens" to embed extra information to help find relevant documents. The fine-tuning data is compiled by de-identifying Wikipedia biographies with a NER model and using Wikipedia pages that are *not* biographies as the background knowledge.

Infilling

After fine-tuning the dense retriever, we create a dataset consisting of Wikipedia biography sanitized chunks and their top ColBERT retrieved text. We use this to train our re-identifier models (both the GLM and PHi-1.5 model). We train them for 2500 steps with a batch size of 64, where each data point is distinct (i.e. there is no repeated training sample). We then use these ColBERT and re-identifier models for the rest of our experiments (for both the Wikipedia biographies and TAB datasets). All models are trained with a single GPU (A100 for Re-identifier, and RTX3090 for ColBERT). In total, the training took 28 hours (9 hours for the GLM, 17 hours for Phi-1.5, and 2 hours for ColBERT) with an additional 50 hours of additional experimenting. In our experiments, the number of passages included for the infilling was set to 1 or 2. When using an autoregressive model such as Phi 1.5, the passage comes before the local context while it comes after the local context for the GLM model.

Dataset	General	All but not original	All
Wikipedia	$41.1^{\pm 18.2}$	$43.7^{\pm 19.2}$	$92.9^{\pm 22.4}$
TAB	$20.7^{\pm 14.1}$	$28.3^{\pm 19.2}$	$100^{\pm 0.0}$

Table 1: Percentage of sanitized spans in a document found in the top-10 retrieved documents using BM25.

4.3 Metrics

For testing, we respectively use 1000 held-out Wikipedia biographies and 300 held-out court cases from the TAB corpus.

We first analyse the performance of the sparse and dense retrievers, and then evaluate the end-to-end performance of the complete system.

Sparse Retrieval To evaluate the performance of the sparse retrieval mechanism, we look at the percentage of masked spans in a sanitized text that can be found in the top 10 retrieved documents.

Dense Retrieval We use both Mean Reciprocal Rank (MRR) and accuracy@k (specifically @1, 5, and 10) to assess the dense retrieval accuracy. If the retrieved text has the span to re-identify, it is considered a positive instance. However, given not all spans have a retrieved chunk with a correct answer, we only look at spans where the masked span exists in one of the retrieved chunks.

Re-identifier We use two metrics to judge the accuracy and performance of our re-identifications. The first is an exact match, where a re-identification is only correct if it outputs the original tokens. The second is token recall where we look at the percentage of tokens in the prediction that are also in the original span. This allows for shorter names that refer to the same person or place (i.e. “President Emmanuel Macron” and “Macron”). In addition to giving results on all tokens, we report results on each NER category/identifier type.

We re-identify the spans in random order, until all spans are replaced. The re-identification is performed “on the fly” – that is, for the re-identification of a masked span, we do a dense retrieval and then use the top results to help re-identify. As spans are re-identified, the dense retrieval has more and more information as previously masked spans are now re-identified.

4.4 Results

4.4.1 Wikipedia Biographies

We first look in Table 1 at the performance of the sparse retrieval. The performance increases along

Knowledge	MRR	Acc@1	Acc@5	Acc@10
WIKIPEDIA				
Not biographies	0.011	0.3	0.9	1.6
All but original	0.018	0.6	1.7	2.9
All	0.303	26.8	33.6	35.8
TAB				
General	0.048	3.1	5.8	8.2
All but original	0.135	8.2	18.8	24.7
All	0.508	43.9	58.4	62.9

Table 2: Performance of the ColBERT model on spans with an existing retrieved chunk from the top-10 retrieved documents by BM25. These results are obtained on fully sanitized texts.

with the level of background knowledge, but we have high variations between biographies (around 20%). This is possibly due to the notoriety of the person in the biography. The more notable a person is, the more likely non-biography texts will contain information on the person.

In addition, we see little difference between Level 2 and Level 3 for this dataset (41.1% vs. 43.7%). Once we include the original biography, we jump to 92.9%. While this difference is high, it is expected, as the document to re-identify is in this setup included as part of the knowledge base.

If we look at Table 2, we see a similar trend where the performance increases as we have more data in the background knowledge. However, the results are relatively low (less than 1% accuracy@1 for non-biographies and all of Wikipedia excluding the original text). Note that as mentioned before, we only consider sanitized spans that have a retrieved chunk with the span from the initial retrieval. Once we include the original text, the performance substantially increases (reaching 26.8% accuracy@1) but is still relatively low, leading us to believe that creating a model adapted to this task could boost performance.

GLM Table 3 provides the exact match results of the end-to-end re-identification performance using the GLM model. We see that similarly to our retriever models, as the background knowledge increases, the re-identification accuracy increases, with a small increase between Level 2 and Level 3 and a big jump once the original text is included in Level 4. We also see that providing retrieved texts improves performance when compared to using just the re-identifier model. While the increase is relatively small overall (1.08% on exact match). When looking at specific categories we can see larger increases such as events, money, and nationalities or political or religious groups (NORP). If we compare using one or two retrieved texts we see a small increase in performance for a relatively large increase in compute time (the sequence length increases by around 40%). Finally, the same trends can be observed in the results for token recall in Table 4. This seems to indicate that the model tends to use alternative versions of the correct span to re-identify. This higher performance holds for almost all NER categories except for CARDINAL, LANGUAGE, NORP, and ORDINAL which tend to have shorter spans. Results in each NER category for token recall can be found in Appendix B.

Phi-1.5 Looking at the Phi-1.5 results in Tables 3 and 4, we see similar trends as our GLM re-identifier model. However, the performance is much worse (not reaching 10% when the original text is included). We posit three possible reasons for this. The first is that the model is less efficient and would require more examples during training to reach good performances. Second, the lack of the right context might mean that the model has a harder time finding the exact nature of the span to re-identify. Finally, the model may have a harder time creating short, non-descriptive answers. More detailed results can be found in Appendix C.

4.4.2 TAB

For the experiments on the TAB dataset, we only used the trained GLM model as it performed better on the Wikipedia biographies experiment. We also used only one retrieved text for re-identification since the gains from using two were minor. Table 1 shows that the retrieval is harder for the court cases in TAB than it was for the Wikipedia biographies. This is probably due to the TAB dataset containing various unique identifiers such as names/codes/dates that do not appear in other cases/reports. However, once we add the original text to the retrieval database, BM25 always finds at least one text with the correct span. This could indicate that the wording used in the court case is unique enough to uniquely identify them (as shown by Weitzenboeck et al. (2022)).

Table 2 details the results of the dense retrieval. We see that the performance of the ColBERT model trained on Wikipedia biographies performs better on the TAB dataset than on the Wikipedia biographies at all levels of background knowledge. This could again be due to the structured style of writing found in court cases. As for Wikipedia biographies,

NER Category	No retrieval	Not Biographies		All but not original		All	
		k=1	k=2	k=1	k=2	k=1	k=2
GLM	**6.83**	**7.91**	**8.13**	**9.10**	**9.37**	**76.42**	**79.47**
CARDINAL	29.55	31.09	31.10	32.01	32.30	85.27	88.87
DATE	3.83	4.35	4.57	5.27	5.43	73.31	77.55
EVENT	9.69	11.60	11.50	14.33	14.41	76.52	78.41
FAC	0.90	1.91	1.68	3.03	3.15	74.38	81.08
GPE	5.62	6.74	7.05	9.00	9.19	78.94	81.86
LANGUAGE	27.80	26.91	26.01	32.29	29.15	84.30	87.89
LAW	5.22	4.35	8.70	9.57	12.17	81.74	84.35
LOC	6.49	7.11	6.07	7.32	8.58	78.45	82.01
MONEY	4.26	7.23	7.66	7.23	8.51	78.72	82.55
NORP	19.75	21.98	22.23	24.67	26.40	84.70	87.28
ORDINAL	49.12	49.93	50.88	50.62	50.26	87.96	88.62
ORG	4.10	5.63	5.72	6.49	7.19	73.62	76.74
PERCENT	2.35	3.53	5.88	3.53	5.88	84.71	80.00
PERSON	0.70	1.64	1.87	2.43	2.40	76.31	79.17
PRODUCT	0.98	2.46	2.94	3.92	5.39	73.53	80.88
QUANTITY	2.75	5.10	3.53	7.45	7.06	80.39	76.47
TIME	5.62	5.62	7.30	6.18	5.06	75.28	76.27
WORK_OF_ART	2.56	3.44	3.75	4.38	4.46	68.85	72.68
PHI-1.5	**0.33**	**0.61**	**0.80**	**0.71**	**0.95**	**9.52**	**8.66**

Table 3: Exact Match of the GLM re-identifier and the Phi-1.5 re-identifier (only overall) at multiple background knowledge levels and number of retrieval texts on the Wikipedia biographies. The overall results are on the same lines as the model name and are bolded. Description of categories can be found in Appendix A.1.

Model	No retrieval	Not Biographies		All but not original		All	
		k=1	k=2	k=1	k=2	k=1	k=2
GLM	13.45	14.73	14.99	16.17	16.53	79.76	82.94
PHI-1.5	1.61	1.78	2.05	1.91	2.24	13.67	11.34

Table 4: Overall token recall of the GLM re-identifier and the Phi-1.5 re-identifier at multiple background knowledge levels and number of retrieval texts on the Wikipedia biographies. More detailed results in Appendices B and C.

both the accuracy and MRR increase along with the levels of background knowledge, reaching up for Level 4 to 43.9% of masked spans appearing in the top document retrieved by our ColBERT model.

Finally, Table 5 provides the results of the end-to-end re-identification with the GLM model. As for the Wikipedia biographies, we see that using any level of background knowledge is beneficial. The benefits of including background knowledge are here slightly more pronounced, with an increase in exact match of 1.16%, 3.18%, and 63.68% between the 4 levels. Similar trends can be found for token recall (with increases of 1.84, 6.37, and 63.59 respectively). If we look at direct identifiers we see that only once we add other TAB cases does our model start re-identifying them (0.53% exact match). Once we include the original court case in the background knowledge (Level 4), the exact match jumps to 59.39%. This is expected given the high performance of the dense retriever for this Level. On the side of the quasi-identifiers, we have the same trend of increasing performance as the background knowledge increases. When looking at specific categories, we see that demographic and location spans are the easiest to identify while code (e.g. case ID) is the hardest. Given that each case has a unique code but that multiple cases can involve people with the same background at the same places this follows quite well.

5 Discussion

Overall, we observe that having background knowledge closely related to the text or spans to re-identify leads to better re-identification of the spans. Usually, unique or uncommon categories of spans (such as direct identifiers) are harder to re-identify than more common ones (such as location, numbers, or demographics). We also saw that using the top retrieved document gives a big performance boost while adding a second retrieved document only minorly improves the performance at the cost of performance.

Surprisingly, the results show that not re-training the models on a new domain still leads to good re-

Entity Category	No retrieval	General Knowledge	All but not original	All
CODE	0.00 / 21.16	0.00 / 10.19	0.31 / 21.18	39.70 / 63.84
DATETIME	0.37 / 5.87	1.53 / 8.10	3.33 / 11.07	77.64 / 81.33
DEM	8.58 / 12.91	6.44 / 11.88	20.81 / 27.70	69.15 / 77.53
LOC	8.31 / 9.55	9.14 / 11.30	10.99 / 12.25	79.55 / 85.38
MISC	0.00 / 9.84	1.12 / 12.96	8.55 / 31.12	47.48 / 77.16
ORG	0.35 / 8.11	5.91 / 16.11	7.96 / 21.45	72.11 / 86.78
PERSON	0.54 / 7.96	0.43 / 8.46	6.66 / 19.00	58.09 / 76.30
QUANTITY	0.69 / 13.95	0.69 / 16.78	1.32 / 19.38	61.38 / 80.14
DIRECT	0.00 / 9.44	0.00 / 5.11	0.53 / 12.32	59.39 / 77.15
QUASI	1.42 / 7.83	2.65 / 10.28	6.05 / 16.60	70.08 / 80.05
ALL	1.32 / 7.98	2.46 / 9.82	5.64 / 16.19	69.32 / 79.78

Table 5: Results of the GLM re-identifier at multiple background knowledge levels and number of retrieval texts on TAB. The first result represents exact match performance and the second is token recall. Description of categories can be found in Appendix A.2.

sults and similar behaviours when deepening background knowledge. This is encouraging since it means we can use data that does not carry privacy risk to create a model to re-identify spans and ultimately improve our sanitization results.

Finally, we also observed that using models that were originally designed for question-answering style retrieval still yielded results when adapting to a slightly different task of finding the most useful documents to re-identify spans (especially at the beginning when no spans are available). However, many improvements are possible and creating a specialized model for this task could be beneficial. This is especially apparent in the results when the original text is in the background knowledge. We also noticed that having both left and right context for the re-identification (GLM) was better than only considering the left context (Phi-1.5), however, this could also be from fine-tuning being less efficient for causal-only models.

6 Conclusion

This paper presented a novel approach to the task of *re-identifying* text documents that had previously been de-identified by masking personal identifiers. Automated re-identification models constitute an important tool to enhancing the robustness of text de-identification methods, and in particular to establish whether the content of a masked text span can be inferred from the context and available background knowledge.

The presented method relies on a retrieval-augmented architecture that comprises a sparse retriever, a dense retriever, and an infilling model fine-tuned to take advantage of the passages extracted in the retrieval phase. The method is evaluated using two datasets, Wikipedia and the Text Anonymization Benchmark (TAB), and with 4 levels of background knowledge. We observed that texts that have been de-identified either through NER (in the case of Wikipedia biographies) or manually (in the case of TAB) can be at least partly re-identified. However, the re-identification performance is strongly dependent on the level of background knowledge which we assume will be available to an adversary. Furthermore, even at the most basic level of background knowledge, some spans are re-identified, although direct identifiers remain relatively safe.

This paper is just a preliminary study on re-identification, in the future, we hope to explore various different angles. Currently, we work with a rather naive dataset of positive/helpful documents where a document is helpful if and only if it contains the span to be re-identified. However, this has several shortcomings such as missing alternative forms of the same span (such as a text containing "President Lincoln" instead of "President Abraham Lincoln"), texts containing the correct span but in a very different context (There were *four* objects considered vs. he had won *four* gold medals), or texts that do not contain the correct span but could still help re-identify the span. Creating a gold dataset manually or using bootstrapping to generate a silver one, could lead to better performance.

We also only focus on re-identifying spans but do not look at how doing so affects the re-identification of who the text is about. Finally, having more granular and grounded levels of background knowledge and defining an order to re-identifying spans could lead to a more detailed understanding of the re-identification task.

Limitations

We only looked at texts in the English language and only used text data to help the re-identification, it is possible that using other types of data such as tables or knowledge graphs could be more helpful to this task. Also, we worked with relatively small models (335M and 1.5B parameters) this is due to some computing constraints and also because it was an initial study into the task. Using larger models with the possibility of In-context learning, could lead to different conclusion on the efficacy of autoregressive models. In addition, both datasets originate from text documents which are otherwise available on the web in clear text. This means that there is a possibility that some of the data has been leaked to the model during the pre-training of it. Using a text which does not have a publicly unsanitized version available could lead to worse performance and even no re-identifying.

Ethical Statement

We acknowledge that creating models to re-identify sanitized texts could help attackers re-identify private data. However, our goal with this paper is to show that if it is possible to re-identify automatically with such models, then using them during sanitization could lead to more robust and future-proof sanitization. One could use these models during sanitization to verify whether certain documents being leaked/released could lead to a higher risk of private data being re-identified.

References

Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. 2022. Efficient training of language models to fill in the middle. *arXiv preprint arXiv:2207.14255*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens.

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*, 24(3):596–606.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Mark Elliot, Elaine Mackey, Kieron O'Hara, and Caroline Tudor. 2016. *The anonymisation decision-making framework*. UKAN Manchester.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

GDPR. 2016. General Data Protection Regulation. European Union Regulation 2016/679.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

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.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research*

and development in Information Retrieval, pages 39–48.

Satyapriya Krishna, Rahul Gupta, and Christophe Dupuy. 2021. ADePT: Auto-encoder based differentially private text transformation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2435–2439, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: **phi-1.5** technical report. *arXiv preprint arXiv:2309.05463*.

Pierre Lison, Ildikó Pilán, David Sanchez, 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, et al. 2023. DeID-GPT: Zero-shot medical text de-identification by GPT-4. *arXiv preprint arXiv:2303.11032*.

Benet Manzanares-Salor, David Sánchez, and Pierre Lison. 2022. Automatic evaluation of disclosure risks of text anonymization methods. In *Privacy in Statistical Databases: International Conference, PSD 2022, Paris, France, September 21–23, 2022, Proceedings*, page 157–171, Berlin, Heidelberg. Springer-Verlag.

Maximilian Mozes and Bennett Kleinberg. 2021. No intruder, no validity: Evaluation criteria for privacy-preserving text anonymization. *arXiv preprint arXiv:2103.09263*.

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.

Anthi Papadopoulou, Yunhao Yu, Pierre Lison, and Lilja Øvrelid. 2022. Neural text sanitization with explicit measures of privacy risk. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 217–229, Online only. 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.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Martin Scaiano, Grant Middleton, Luk Arbuckle, Varada Kolhatkar, Liam Peyton, Moira Dowling, Debbie S Gipson, and Khaled El Emam. 2016. A unified framework for evaluating the risk of re-identification of text de-identification tools. *Journal of biomedical informatics*, 63:174–183.

Viju Sudhi, Sinchana Ramakanth Bhat, Max Rudat, and Roman Teucher. 2024. Rag-ex: A generic framework for explaining retrieval augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2776–2780.

Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the scrub system. In *Proceedings of the AMIA annual fall symposium*, pages 333–337. American Medical Informatics Association.

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.

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.

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, Tokyo, Japan. Association for Computational Linguistics.

Wanrong Zhu, Zhiting Hu, and Eric Xing. 2019. Text infilling. *arXiv preprint arXiv:1901.00158*.

A Description of NER Categories

A.1 Wikipeadia Biographies

These description come directly from Spacy.[4]

CARDINAL Numerals that do not fall under another type

DATE Absolute or relative dates or periods

EVENT Named hurricanes, battles, wars, sports events, etc.

FAC Buildings, airports, highways, bridges, etc.

GPE Countries, cities, states

LANGUAGE Any named language

LAW Named documents made into laws.

LOC Non-GPE locations, mountain ranges, bodies of water

MONEY Monetary values, including unit

NORP Nationalities or religious or political groups

ORDINAL "first", "second", etc.

ORG Companies, agencies, institutions, etc.

PERCENT Percentage, including "%"

PERSON People, including fictional

PRODUCT Objects, vehicles, foods, etc. (not services)

QUANTITY Measurements, as of weight or distance

TIME Times smaller than a day

WORK_OF_ART Titles of books, songs, etc.

A.2 TAB

These descriptions come from the paper (Pilán et al., 2022).

CODE Numbers and identification codes, such as social security numbers, phone numbers, passport numbers, or license plates.

DATETIME Description of a specific date, time, or duration.

DEM Demographic attributes of a person, such as native language, descent, heritage, ethnicity, job titles, ranks, education, physical descriptions, diagnosis, birthmarks, ages.

LOC Places and locations, such as cities, areas, countries, addresses, named infrastructures, etc.

MISC Every other type of personal information associated (directly or indirectly) to an individual and that does not belong to the categories above.

ORG Names of organizations, such as public and private companies, schools, universities, public institutions, prisons, healthcare institutions, non-governmental organizations, churches, etc.

PERSON Names of people, including nicknames/aliases, usernames, and initials.

QUANTITY Description of a meaningful quantity, e.g., percentages or monetary values.

[4] https://spacy.io/

B Token recall results for the GLM

Table 6 contains the detailed token recall results of the GLM re-identifier.

NER Category	No retrieval	Not Biographies		All but not original		All	
		k=1	k=2	k=1	k=2	k=1	k=2
GLM	**13.45**	**14.73**	**14.99**	**16.17**	**16.53**	**79.76**	**82.94**
CARDINAL	28.75	30.74	30.93	31.71	32.15	83.94	87.52
DATE	19.27	20.72	21.04	22.29	22.84	77.43	80.93
EVENT	31.63	34.11	34.01	36.45	36.15	86.96	89.69
FAC	11.08	11.73	11.75	13.89	13.81	80.63	86.67
GPE	6.41	7.30	7.56	9.79	10.04	79.21	82.74
LANGUAGE	27.21	23.81	25.09	33.45	30.55	83.73	89.94
LAW	32.82	34.10	39.47	35.19	40.05	91.34	94.55
LOC	10.74	11.02	10.56	10.34	12.45	82.12	84.43
MONEY	24.94	26.07	27.97	29.10	29.68	87.24	91.26
NORP	17.16	19.62	20.12	23.08	24.67	84.69	87.45
ORDINAL	43.99	44.54	46.27	46.97	45.94	84.81	86.17
ORG	19.08	20.63	21.02	21.73	22.29	80.37	83.71
PERCENT	23.86	29.52	28.14	31.34	32.27	92.21	90.99
PERSON	3.16	4.20	4.36	5.29	5.35	79.07	81.76
PRODUCT	8.04	10.42	9.12	10.72	12.27	79.94	86.20
QUANTITY	27.00	29.67	28.10	34.67	37.96	90.16	88.31
TIME	21.28	23.84	22.02	22.75	19.92	82.17	85.25
WORK_OF_ART	12.50	13.35	13.81	14.81	14.99	77.89	81.34

Table 6: Token recall of the GLM re-identifier at multiple background knowledge levels and number of retrieval texts on the Wikipedia biographies. The overall results are on the same lines as the model name and are bolded. Description of categories can be found in Appendix A.1.

C Phi-1.5 Results

Table 7 contains the detailed exact match results of our Phi-1.5 re-identifier model. While Table 8 contains the detailed token recall results.

NER Category	No retrieval	Not Biographies		All but not original		All	
		k=1	k=2	k=1	k=2	k=1	k=2
Phi-1.5	**0.33**	**0.61**	**0.80**	**0.71**	**0.95**	**9.52**	**8.66**
CARDINAL	0.86	0.74	1.19	0.74	1.12	7.82	9.29
DATE	0.42	0.40	0.57	0.52	0.58	6.46	7.06
EVENT	0.42	0.87	1.50	1.18	1.90	8.85	9.48
FAC	0.17	0.22	0.22	0.34	0.22	7.65	9.67
GPE	0.96	1.20	1.57	1.27	1.75	8.50	9.61
LANGUAGE	0.00	0.46	0.46	0.93	0.93	14.81	11.57
LAW	0.00	0.00	0.00	0.00	0.00	2.63	11.40
LOC	0.00	0.21	0.21	0.42	0.21	6.69	11.92
MONEY	3.98	2.98	3.40	3.40	4.68	27.66	23.40
NORP	0.52	1.27	1.31	1.68	2.36	12.94	13.06
ORDINAL	3.77	2.39	4.41	2.20	4.17	8.14	10.39
ORG	0.56	0.61	0.70	0.73	1.00	7.71	7.27
PERCENT	0.00	0.00	0.00	0.00	0.00	5.88	11.76
PERSON	0.12	0.23	0.19	0.28	0.25	11.06	7.26
PRODUCT	0.00	0.00	0.00	1.47	0.00	11.76	9.80
QUANTITY	0.00	0.00	0.00	0.00	0.00	3.53	3.92
TIME	0.00	0.00	0.00	0.56	0.00	1.69	6.18
WORK_OF_ART	0.04	0.25	0.28	0.30	0.47	21.07	15.97

Table 7: Exact match of the Phi-1.5 re-identifier at multiple background knowledge levels and number of retrieval texts on the Wikipedia biographies. The overall results are on the same lines as the model name and are bolded. Description of categories can be found in Appendix A.1.

NER Category	No retrieval	Not Biographies		All but not original		All	
		k=1	k=2	k=1	k=2	k=1	k=2
Phi-1.5	**1.61**	**1.78**	**2.05**	**1.91**	**2.24**	**13.67**	**11.34**
CARDINAL	0.60	0.49	0.74	0.46	0.76	5.00	5.73
DATE	1.25	1.25	1.81	1.56	1.83	8.16	10.28
EVENT	3.43	3.70	4.93	4.00	5.80	18.71	18.21
FAC	1.26	1.42	1.62	1.33	1.53	16.04	15.48
GPE	0.82	1.04	1.18	1.12	1.42	9.64	8.13
LANGUAGE	0.00	0.16	0.16	0.81	0.78	11.85	6.11
LAW	3.24	2.51	3.33	1.76	3.00	17.98	23.47
LOC	0.59	0.47	0.65	0.46	0.49	10.52	11.51
MONEY	12.59	15.14	14.06	13.84	13.82	40.11	35.02
NORP	0.54	0.84	0.99	0.96	1.55	10.38	7.99
ORDINAL	1.77	0.98	1.69	0.95	1.66	4.94	5.65
ORG	3.03	3.00	3.57	3.20	3.89	15.58	13.49
PERCENT	0.00	0.49	0.93	1.42	0.49	12.90	15.13
PERSON	0.55	0.82	0.74	0.86	0.88	13.48	8.87
PRODUCT	1.64	2.11	1.03	2.63	1.14	18.07	14.40
QUANTITY	0.70	2.53	2.96	4.30	5.05	15.02	10.71
TIME	1.32	3.52	2.37	2.41	2.12	9.79	15.20
WORK_OF_ART	5.16	5.46	5.77	5.51	5.82	31.86	24.94

Table 8: Token recall of the Phi-1.5 re-identifier at multiple background knowledge levels and number of retrieval texts on the Wikipedia biographies. The overall results are on the same lines as the model name and are bolded. Description of categories can be found in Appendix A.1.