Multi-Token Completion for Text Anonymization

Pulkit Madaan
Johns Hopkins University
pmadaan2@jhu.edu

Krithika Ramesh
Johns Hopkins University
kramesh3@jhu.edu

Lisa Bauer
Amazon
bauerlb@amazon.com

Charith Peris
Amazon
perisc@amazon.com

Anjalie Field
Johns Hopkins University
anjalief@jhu.edu

Abstract

Text anonymization is a critical task for enabling research and development in high-stakes domains containing private data, like medicine, law, and social services. While much research has focused on redacting sensitive content from text, substantially less work has focused on what to replace redacted content with, which can enhance privacy and becomes increasingly important with greater levels of redaction. In this work, we formulate predicting replacements for sensitive spans as a research task with principled use-inspired evaluation criteria. We further propose a multi-token completion method for accomplishing this task that is designed to preserve consistency with low compute requirements, thus facilitating practitioners to anonymize data locally before sharing it externally. Human and automated annotations demonstrate that our approach produces more realistic text and better preserves utility than alternative infilling methods and differentially private mechanisms across multiple domains without retraining. Overall, our work explores the under-studied task of what to replace redacted content with and contributes grounded evaluations capturing utility, facilitating future work.

1 Introduction

Advances in NLP have been accelerated by access to large datasets for model pre-training and publicly available benchmarks for standardized evaluation. Similar advances could be immensely useful in high-stakes domains with sensitive text data like law (Zhong et al., 2020), medicine (Panchbhai and Pathak, 2022), and social services (Gandhi et al., 2023). Realizing this potential, however, requires sharing private data with researchers and practitioners in order to train and evaluate models on domain-specific tasks and terminology. While data use agreements can facilitate limited sharing (Gandhi et al., 2023; Field et al., 2023), they only enable access for a small group of people, and they still carry risks of misuse or accidental leakage.

Anonymizing or sanitizing text offers a safer way to enable broader access. Painstakingly careful efforts to remove direct identifiers in compliance with policies like HIPAA have enabled increased access to healthcare data (Johnson et al., 2016; Goldberger et al., 2000), and numerous tools and shared tasks have focused on detecting direct identifiers like names and social security numbers from text (Lison et al., 2021; Mendels et al., 2018; Stubbs and Uzuner, 2015; Stubbs et al., 2017). These methods, however, fail to achieve true anonymization. Experts speculate that no method is likely to be able to guarantee 100% recall, (Carrell et al., 2012), and further, removing *direct* identifiers is insufficient. True anonymization requires additionally removing *quasi-identifiers* like nationality or physical descriptions that may lead to re-identification, especially when combined with external data sources (Pilán et al., 2022).

A potential solution lies in replacing identifiers with realistic alternatives rather than simply redacting them (Hirschman and Aberdeen, 2010; Carrell et al., 2012). This approach, drawing from the theory of Hiding In Plain Sight (HIPS) and backed with initial empirical evidence (Carrell et al., 2012), proposes that replacing the detected identifiers with realistic surrogates makes it more difficult to distinguish any leaked real identifiers. Replacement also offers a way to support increased levels of redaction, such as the redaction of quasi-identifiers, without compromising text quality. Despite these motivations, what to replace identifiers with is a surprisingly under-studied problem; in their survey of text anonymization, Lison et al. (2021) describe it as "rarely addressed in NLP", and even recently-proposed LLM-based methods for redaction still assume rule-driven entity typing systems for replacement (Ji et al., 2025).

In this work, we propose approaching this task

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5894–5908

through directly predicting spans of text to replace redacted content, and we develop a lightweight infilling approach that uses masked-language models to produce realistic context-aware alternatives. Masked-language models like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and Longformer (Beltagy et al., 2020) are ideal for the task of generating replacements because their training objective specifically targets predicting replacements for masked tokens. They also require minimal compute compared to LLMs, thus allowing practitioners to easily anonymize data locally. However, off-the-shelf, these models are only designed to generate a single token from a limited vocabulary. Instead, we build on a multi-token completion method (Kalinsky et al., 2023), in which we augment model vocabulary with multi-token phrases to facilitate generating fluent replacements for redacted content.

We use the text anonymization benchmark (TAB) (Pilán et al., 2022) to evaluate our model in an end-to-end pipeline, where we first train and deploy a supervised model to identify direct and quasi-identifiers in text, and then use our infilling model to replace the redacted content. We assess the full system through two human annotation tasks: plausibility of proposed replacements and successful anonymization. We further evaluate the system for our proposed use case: training NLP models and thus enabling future development in domains with sensitive data. Finally, we assess the generalizability to healthcare data using clinical notes from the i2b2/VA Shared-Task (Uzuner et al., 2012) without retraining the detection or the infilling model, making our model truly compute minimal with high potential for achieving local anonymization in real-world use cases.

Our contributions include (1) the formulation of predicting replacement spans for sensitive information as an NLP task with principled evaluation criteria grounded in realistic use cases; (2) a specific method for predicting replacements, that leverages light-weight masked-language-models in order to support local anonymization;[1] (3) novel evaluation criteria that captures the lack in usability of existing anonymization schemes; (4) empirical results demonstrating the effectiveness of our method over data from two domains, highlighting better utility preservation against just redaction and differential privacy methods. Our results demonstrate that the need for efficient and realistic replacements has grown in importance as use cases for text data have become increasingly complex, thus encouraging future work in this area.

[1]Code and data will be provided upon publication

2 Methodology

Given a text consisting of tokens $t_1, ..., t_n$, we assume we have a sensitive span detection model, which denotes, for example, that $t_i, ..., t_j$ contains information that needs to be redacted. This sensitive span detection model can vary depending on the level of anonymization needed. Our task is to generate tokens to replace $t_i, ..., t_j$, where tokens should be both grammatical and realistic enough to preserve the general meaning of the text. The new text with replaced tokens is what we call *pseudoanonymized* going forward. As the goal of this work is anonymizing sensitive text that cannot be shared externally, the replacement system needs to be lightweight enough to run locally on a CPU or small GPU, which is easily accessible to practitioners. This goal also precludes the use of API-based LLMs.

Our primary approach draws from the observation that pretrained masked language models (MLMs) are uniquely suited to this task: their training objective explicitly targets predicting replacements for masked tokens. MLMs are also smaller and less compute-intensive, making them more feasible for data stewards like hospitals or courts to run locally. However, off-the-shelf, these models are only capable of replacing masked tokens with a single (subword) token that is best suited based on the context. Single token replacement is highly restrictive for our use case, as sensitive spans like names or addresses often include multiple words, and the majority of text entities of any type are usually multiple tokens long (Kalinsky et al., 2023). We use a straightforward adaptation of the task-agnostic method for multi-token completion proposed in Kalinsky et al. (2023).

Multi-Token Completion (MTC) The overall approach involves augmenting the MLM's vocabulary with multi-token phrases. Then, the model is fine-tuned over sentences containing these phrases, where only the output embedding matrix is updated. This approach expands only the decoder's embedding matrix with new multi-token words, which minimizes the required compute and additional parameters.

In Kalinsky et al. (2023)'s task-agnostic set-up,

they identify multi-token phrases to add to the vocabulary by extracting common noun phrases from Wikipedia. Out of the collected noun phrases, only 10.2% are comprised of a single token, further motivating the need for multi-token completion. Kalinsky et al. (2023) show that their MTC approach outperforms alternatives, such as T5-3B-like models (Raffel et al., 2020), and is significantly more efficient in terms of parameter count. We compare performance using the same set of Wikipedia noun phrases with our proposed adaptation that more directly targets sensitive span replacement.

Sensitive-Span Multi-Token Completion (SS-MTC) Kalinsky et al. (2023) extract common multi-token noun phrases for vocabulary expansion, which is sufficient for showing the benefits of their approach in a task-agnostic way. However, sensitive spans are often not noun phrases, suggesting that we can improve performance through more tailored vocabulary selection. Thus, we build a corpus of candidate replacement spans that are more similar to actual sensitive spans by running a sensitive span detection model over public data (Wikipedia). We then expand the decoder vocabulary of the pretrained MLM using this set of spans and fine-tune the output embedding matrix on extracted sentences similar to MTC. This model is only required for inference for the task of anonymization, reducing the compute load significantly to a matter of hours on a single T4 GPU.

Naive Static-3 (NS3) In addition to the original MTC approach and our proposed adaption, we also compare against two simpler methods for enabling multi-token completion. In the first, we replace every marked redaction with 3 consecutive mask tokens. This method effectively forces the MLM to infill with 3 token-long entities.

Naive Adaptive (NAd) Finally, instead of replacing with a fixed number of mask tokens, we replace every token in the original entity with a mask token. This approach enables the infilled entity to not be fixed to a particular length and to be the same as the original in token length.

3 Experiments

We evaluate each infilling method as a component of a full end-to-end pipeline for text anonymization. The first half of the pipeline is the detector. This model identifies the sensitive spans in the document. The second half is the infilling model, which uses the context around the sensitive span to provide a replacement.

3.1 Data

Our primary data is the Text Anonymization Benchmark (TAB) (Pilán et al., 2022). This dataset contains 1,268 court case documents in English from the European Court of Human Rights. The dataset is split across train, dev, and test with 1014, 127, and 127 documents, respectively, and has an average document length of 16233 tokens.

Unlike other text de-identification datasets, TAB was designed specifically for *anonymization*: each document was annotated for both direct and quasi-identifiers by multiple annotators. Quasi-identifiers make for the majority of the marked sensitive spans at 93.47%. For this reason, we use the TAB dataset to train the detector model, both for identifying Wikipedia spans to add to the MLM vocabulary and to identify sensitive spans to replace in evaluation data. We additionally use the LexAbSumm dataset (T.y.s.s. et al., 2024) and clinical notes from the i2b2/VA Shared-Task and Workshop (Uzuner et al., 2012) in order to evaluate downstream utility tasks and cross-domain generalizability, which we discuss in more detail in §3.3.

3.2 Models

We use long context models for both parts of the pipeline as splitting as a strategy would be detrimental to anonymization robbing the model of long-range context. We specifically use the base Longformer model , *longformer-base-4096*, (Beltagy et al., 2020) as it can accommodate 4096 tokens at a time, which is larger than most of the documents (93.3%) in our experiments.

Sensitive span detection The first part of the pipeline is a binary token-level prediction task, where each token is labeled as sensitive or not sensitive. We fine-tune Longformer on the TAB dataset for token classification, using the same settings detailed in Pilán et al. (2022). Since the Longformer tokenizes the text on a sub-word level, we combine consecutively predicted sensitive tokens into a single entity and mark it for redaction.

To confirm that our sensitive span detection model performs accurately enough for use in our evaluation pipeline, we evaluate the detector on the held-out TAB test set (Table 1), reporting recall and precision at the token and mention level as described in Pilán et al. (2022). Our model achieves

	Recall		Precision	
	Tok.	**Men.**	**Tok.**	**Men.**
Pres.	0.766	0.713	0.582	0.438
Long.	**0.929**	**0.905**	**0.882**	**0.743**

Table 1: Precision and Recall evaluated on the TAB dataset's test split at Token and Mention granularity. Our detector model (Longformer) achieves strong performance, especially in comparison to Microsoft's off-the-shelf anonymizer (Presidio).

strong performance, especially in comparison to an off-the-shelf anonymization model (Microsoft's Presidio[2]).

Infilling Models For all infilling approaches, we similarly use Longformer as the starting MLM. When conducting infilling, we combine all the positively predicted consecutive sensitive tokens from the detector and treat them as a single entity for replacement.

For MTC, we use the same set of extracted noun phrases and fine-tuning data as Kalinsky et al. (2023). This setup entails expanding Longformer's vocabulary with 93K phrases that occur at least 500 times on Wikipedia and fine-tuning the model on 50 unique sentences for each added phrase. We conduct fine-tuning for the multi-token completion objective for 10 epochs using the Extended-Matrix (EMAT) decoder approach from Kalinsky et al. (2023).

For SS-MTC, we use our detector model trained on TAB to select sensitive spans from Wikipedia data. For each span with a frequency greater than 500, we select 50 Wikipedia sentences containing the span for fine-tuning, mimicking the MTC setup. Our model identifies 135k unique spans making the final corpus 1.5x the size of the original noun phrase corpus constructed by Kalinsky et al. (2023). For training, we divide the extracted sentences containing each span into train (90%), validation (5%), and test (5%) sets. Given the larger vocabulary size, we then fine-tune Longformer on the training set using the same EMAT approach for 20 epochs.

3.3 Evaluation

Human Evaluation We conduct human evaluation studies to compare the performance of each infilling approach. We designed two tasks: (1) we ask the annotators to rate the best and the worst approach in terms of language quality and (2) we verify that the replacements from all four approaches do not leak information about the original entity.

As the TAB dataset contains long documents, for easier annotation we divided the data into smaller chunks (delimited by newline characters) and selected 340 sentences containing replacement spans for annotating. We divided these sentences into 17 buckets of 20 questions each. 15 of these buckets (300 sentences) were used for quality annotations (task 1), which we conducted using Prolific (details on instructions and questions in §B). Annotators were screened to make sure they were fluent in English, from the U.S. or a European Union country, and their occupation role was legal in nature owing to the fact that the data is English court case documents from the European Court. We solicited annotations from 5 different annotators for each bucket of 20 sentences. Every annotator was paid at $20/hour at par with study and pay standards.

We used the remaining 2 buckets (40 sentences) to collect annotations on privacy leakage(task 2). Each sentence was annotated by four authors of this work. We do not report results for a larger-scale study as we found no evidence of privacy leakage for the proposed methods in our internal annotations (reported in §4).

Automated evaluations We construct two automated metrics for comparing infilling approaches. First, we use Spacy's xx_ent_wiki_sm NER model (Honnibal et al., 2020) to infer entity types of the original and replacement spans, and we compute how often the replacement span matches the entity type of the original span. Second, we use language model perplexity to evaluate the consistency between the original text and text generated by language models trained on the pseudoanonymized text. More specifically, we finetune DistilGPT2[3] Language Model (LM), a GPT2 model distilled using the process proposed by Sanh et al. (2019). The LM finetuned on the original data is used to compute the perplexity of the text generated by LMs finetuned on each pseudoanonymized text under each infilling method. Both of these metrics are constructed to capture if the proposed infilling method produces text that is plausible and consistent as compared to the original data.

Downstream Utility Evaluation We further consider an application-inspired use case: if a practitioner uses our proposed methods to anonymize

[2]https://microsoft.github.io/presidio/

[3]https://huggingface.co/distilbert/distilgpt2

text before sharing data with NLP contractors or researchers, would the contractors or researchers be able to share back models that perform well on real (not anonymized) text? To evaluate this setup, we train NLP models on text pseudoanonymized using each infilling approach and evaluate their performance over held-out original data. We use two NLP tasks: summarization, which is a popular task in NLP for legal documents, and language modeling, which has become commonplace for pretraining in-domain models prior to task-specific fine-tuning. For summarization, we finetune the LongT5 tglobal-base model (Guo et al., 2022) on the LexAbSumm dataset (T.y.s.s. et al., 2024), where we pseudoanonymize the training set and evaluate over the original test set. For language modeling, we use a similar setup as automated evaluations, but we finetune the LM over the pseudoanonymized text and evaluate the perplexity of the held-out original text with these models, rather than the inverse.

Domain Generalizability Finally, in order to evaluate domain generalizability, we additionally use a dataset of clinical notes from the MIMIC-II Database annotated for coreference as a part of the i2b2/VA Shared-Task and Workshop in 2011 (Uzuner et al., 2012). In this setting, we use the exact same detection and infilling models developed over TAB without any training over the i2b2 data. We evaluate downstream utility for coreference resolution using the SynthTextEval[4] pipeline. First, we use a s2e coreference resolution model (Kirstain et al., 2021) fine-tuned over the i2b2 training set to infer "silver annotations" over the pseudoanonymized clinical notes. This process simulates how a researcher or practitioner might manually annotate the data for model development. Then, we fine-tune the coreference resolution model over the silver-annotated data and evaluate it over the real annotations to assess if the pseudoanonymized data is of sufficient quality to support model training.

4 Results

4.1 Human Evaluations

Table 2 reports results from Task 1, the human evaluations of each infilling approach, where annotators were instructed to select the best and worst model output in terms of grammatical correctness and plausibility of meaning. Interannotator agreement over the exact choice of best (Cohen's $\kappa = 0.20$) and worst (Cohen's $\kappa = 0.25$) is higher than chance, but low overall as methods sometimes had similar outputs (for example, both naive methods might output similar text), and we do not direct annotators on which option to select in that case.[5] We nevertheless can validate our annotations by comparing how often annotators had extreme disagreements. If we consider two annotators to disagree only if one of them selected a model as the best when the other selected it as the worst, we end up with an agreement score of 0.75. Thus, annotators rarely conflated best and worst models, and our agreement overall is high.

SS-MTC performs the best across all metrics. Both the MTC approaches easily outperform the naive approaches, and annotators selected SS-MTC as the best most often (in 29.7% of the samples) and the worst least often (11.7% of the samples). Thus, multi-token completion, in general, outputs more plausible replacements than naive approaches, and tailoring the vocabulary of multi-token phrases to sensitive span detection outperforms a vocabulary

	Mode (%)		Overall (%)	
	Best	**Worst**	**Best**	**Worst**
NS3	20.7	30.3	19.6	26.2
NAd	21.3	37.0	19.2	33.8
MTC	28.3	21.0	27.2	22.2
SS-MTC	**29.7**	**11.7**	**34.0**	**17.8**

Table 2: For each question, annotators were shown replacements proposed by each model and instructed to select the best and worst in terms of fluency and overall language quality. For each model, we report the percent of time it was selected as the best or worst by the majority of annotators (Selected Mode). We also report the percent of the time it was selected as the best or worst over all annotators and data points. SS-MTC performs the best by all metrics.

	NS3	**NAd**	**MTC**	**SS-MTC**
% Anon.	100	90	100	100

Table 3: Total percentage of data rated as anonymized (task 2) by human annotators, where anonymization is determined based on the majority vote of annotators.

[4] https://github.com/kr-ramesh/synthtexteval/

[5] We determined through several internal pilot studies that directing annotators which model to select when equivalent would make instructions complicated and confusing, and accepting lower agreement would result in more reliable annotations.

Original	... statutory rate, running from **23 October 1996**, the date on which ...
NS3	... statutory rate, running from **1- date**, the date on which ...
NAd	... statutory rate, running from **1- date**, the date on which ...
MTC	... statutory rate, running from **Michaelmas**, the date on which ...
SS-MTC	... statutory rate, running from **30 June**, the date on which ...
NS3	... were received from the **Government, Justice** and **the of the** Governments ...
NAd	... were received from the **United States** and **the** Governments ...
MTC	... were received from the **petitioners** and **Lockerbie** Governments ...
SS-MTC	... were received from the **Czech** and **Regional** Governments ...

Table 4: Examples of replacing direct or quasi-identifiers in the TAB data set using different in-filling approaches. SS-MTC leads to more fluent and realistic replacements than the naive approaches.

Method	Precision	Recall	F1	Acc.
NS3	0.44	0.35	0.35	0.63
NAd	0.42	0.34	0.34	0.63
MTC	0.38	0.38	0.37	0.61
SS-MTC	**0.45**	**0.42**	**0.42**	**0.66**

Table 5: NER automated evaluation. SS-MTC maintains the most consistency in entity type as compared to the original data.

Method	Perplexity	CI
Original	5.12	(4.93, 5.35)
Masked	12.16	(11.57, 12.82)
NS3	25.90	(23.80, 28.23)
NAd	29.29	(27.06, 32.19)
SanText+	119.48	(109.15, 131.58)
CusText+	78.40	(74.25, 82.82)
CluSanT	26.42	(25.10, 27.97)
LLM	9.57	(9.05, 10.10)
MTC	**7.00**	**(6.71, 7.33)**
SS-MTC	7.16	(6.88, 7.50)

Table 6: Automated evaluation on language modeling task using the LexAbSumm dataset. Perplexity of the text generated by LMs trained on each version of the pseudoanonymized text is evaluated against the LM trained on the Original data. Numbers inside the parentheses indicate 95% confidence interval computed through bootstrapping over the test set.

of generic noun phrases.

It is possible that the improved performance of SS-MTC could come at the cost of anonymization. By using the same sensitive span detection model to select the vocabulary of replacements, SS-MTC could introduce privacy leakage. For example, "New York City" could be labeled as a quasi-identifier in TAB, and as this span also occurs in Wikipedia data, it could be added to the expanded vocabulary. Checking for exact matching between replacements and original spans would not sufficiently evaluate this concern: for example, "New York City" could be replaced with "NYC". Instead, we manually annotate if replacement spans are sufficiently anonymizing (Table 3), with high interannotator agreement (Cohen's Kappa=0.79). All models preserve anonymity to a great extent, and the only identified leakage occurred for NAd, not for either MTC infilling approach.

In Table 4 we show examples demonstrating why SS-MTC replacements are more plausible as compared to naive (NS3, NAd) replacements. SS-MTC replaces spans with similar but non-identical entities, e.g. replacing a date with a different date. In some cases, the model replaces a specific entity with a more generic one, e.g., replacing the name of a specific government with "Regional".

4.2 Automated evaluations

Tables 5 and 6 report results for the NER and LM automated evaluations, respectively. In the NER results (Table 5), MTC surprisingly performs worse than the naive approaches in accuracy, though it does still outperform naive results in F1. NER consistency is an imperfect metric of quality: as demonstrated by the examples in Table 4, a proposed replacement can still retain fluency and plausibility even if the NER type changes. Regardless, this metric does demonstrate that SS-MTC better preserves entity consistency than other approaches.

In Table 6, unsurprisingly all pseudoanonymization methods result in worse (higher) perplexity than original unmodified data. SS-MTC and MTC best reduce this gap, with MTC slightly outperforming SS-MTC on this metric. Both methods far outperform the naive approaches, which increase perplexity even more than just masking out sensi-

Method	Rouge-1	Rouge-2	Rouge-L	Rouge-Lsum
Original	40.48: (39.1, 41.9)	17.05: (15.9, 18.1)	28.06: (26.9, 29.2)	28.08: (27.0, 29.1)
Masked	38.87: (37.5, 40.4)	15.36: (14.3, 16.4)	26.74: (25.7, 27.9)	26.80: (25.7, 27.9)
NS3	38.09: (36.7, 39.4)	15.32: (14.3, 16.5)	26.71: (25.6, 27.9)	26.77: (25.6, 27.8)
NAd	38.88: (37.4, 40.3)	15.36: (14.3, 16.5)	26.80: (25.7, 27.9)	26.80: (25.7, 27.9)
MTC	38.70: (37.4, 40.1)	15.56: (14.5, 16.6)	26.54: (25.5, 27.6)	26.58: (25.5, 27.7)
SS-MTC	**39.23**: **(37.8, 40.6)**	**15.63**: **(14.7, 16.8)**	**27.24**: **(26.2, 28.3)**	**27.17**: **(26.1, 28.4)**

Table 7: Downstream evaluation on the summarization task using the LexAbSumm dataset. SS-MTC performs the best overall, though 95% confidence intervals are overlapping, likely due to the sparsity of sensitive spans in reference summaries.

Method	Perplexity	CI
Original	14.60	(13.90, 15.38)
Masked	16.30	(15.56, 17.15)
NS3	16.26	(15.54, 17.06)
NAd	16.20	(15.46, 17.02)
SanText+	37.57	(36.05, 39.51)
CusText+	28.30	(27.34, 29.32)
CluSanT	**15.61**	(13.66, 18.19)
MTC	16.00	**(15.29, 16.85)**
SS-MTC	16.17	(15.44, 16.89)

Table 8: Downstream Language Modeling evaluation on LexAbSumm dataset. Perplexity of the "Original" text is evaluated against each version's LM. Numbers inside the parentheses indicate 95% confidence interval.

Method	Mention F1	Coref F1
Original	0.800 ± 0.005	0.706 ± 0.006
Masked	0.636 ± 0.132	0.506 ± 0.136
NS3	0.699 ± 0.009	0.577 ± 0.008
NAd	0.703 ± 0.005	0.580 ± 0.006
MTC	0.724 ± 0.024	0.614 ± 0.023,
SS-MTC	**0.726 ± 0.022**	**0.617 ± 0.020**

Table 9: Generalizability Coreference evaluation on i2b2 dataset over multiple seeds. We report metrics for detection of mentions and coreference resolution.

tive spans without replacing them.

Although compute costs preclude LLM and DP-based approaches as practical solutions to this task, for completeness we report performance of the LLM-based "Sampling - Special Token" approach from Dou et al. (2024) and three text DP approaches, SanText, CusText, and CluSanT (Yue et al., 2021; Chen et al., 2023; Awon et al., 2025) in Table 6. During our experiments, we found that LLM responses were not always correctly structured as instructed, leading to incomplete anonymization, which is an additional flaw of the LLM-based approach that is not trivial to correct. All of the methods clearly result in lower text quality, and thus we do not conduct the full suite of evaluation metrics over them.

4.3 Downstream Utility Evaluation

Similar to automated LM evaluations, models trained over pseudoanonymized text perform worse than models trained over original data in the downstream utility evaluations (Tables 7 and 8). MTC and SS-MTC perform consistently well for language modeling (Table 8; CluSanT achieves the lowest mean perplexity but has the largest confidence interval, even encompassing "Original"), and SS-MTC performs the best for summarization (Table 7). However, differences between models are generally small with overlap in 95% confidence intervals, likely reflecting the general sparsity of sensitive spans. This sparsity is especially true for LexAbSumm references summaries, which often omit details in condensing documents into summaries. Overall, while the human and automated evaluations clearly demonstrate the benefits of MTC and SS-MTC, the downstream utility evaluations show similar, though less conclusive, trends.

4.4 Generalizability evaluation

Cross-domain results show similar performance as in-domain results (Table 9). There is a large decline in performance when training models on original data (Coref F1=0.706) as compared to masked data (0.506), which offers strong evidence for the need to replace redacted tokens with realistic values in order to preserve data utility. The two MTC methods best reduce this performance gap, with SS-MTC (0.617) achieving slightly better performance than MTC (0.614). These results indicate high potential for the practical usability of SS-MTC: infilling with this approach without

conducting any domain-specific training is able to substantially reduce the performance gap resulting from de-identification.

5 Related Work

Some methods for redacting sensitive content have used rule-based heuristics (Ruch et al., 2000; Neamatullah et al., 2008), while others have used NER-centric methods (Adams et al., 2019; Hassan et al., 2018; Ribeiro et al., 2023), sometimes with fine-grained classifications of PHI (Sweeney, 1996; Alfalahi et al., 2012; Eder et al., 2019; Chen et al., 2019; Volodina et al., 2020; Mamede et al., 2016). These approaches remain limited as they lack diversity, do not adapt well to various domains and contexts, and are inadequate at capturing the long-range dependencies that can be characteristic of long-form text. As our results show, straight redaction degrades text utility for downstream tasks.

Similar to our work, some prior work has combined token classification strategies with language model infilling. Recent approaches use LLMs with few-shot prompting to replace sensitive spans (Yermilov et al., 2023; Vats et al., 2024; Staab et al., 2025; Pissarra et al., 2024). LLM-based approaches are compute-intensive, making local anonymization costly. API-based models rather than local inference are not a sufficient solution for text anonymization, as they require sending raw private data to a third party, which is exactly what anonymization aims to render unnecessary.

An alternative line of work has incorporated differential privacy (DP) in text anonymization. Unlike methods based on legal standards/policies like HIPAA, these approaches rely on the theoretical guarantee that DP provides to ensure privacy protection. For example, SanText (Yue et al., 2021), CusText (Chen et al., 2023), and CluSanT (Awon et al., 2025) use DP mechanisms to substitute tokens in the input text in a manner that aims to preserve overall content coherence. Other work has proposed generating entirely synthetic data from language models fine-tuned with DP (Yue et al., 2023; Kurakin et al., 2023; Mattern et al., 2022; Putta et al., 2023). However, these approaches incur a high initial cost, and they add substantial noise to the data, even noising non-sensitive content, which results in degraded text utility (Ramesh et al., 2024). Furthermore, despite theoretical privacy guarantees, they have unexpected leakage due to the difficulty of formalizing DP for text, thus failing to satisfy law and policy requirements (Lukas et al., 2023; Ramesh et al., 2024).

6 Discussion

Results from human (§4.1) and automated evaluations (§4.2) demonstrate that MTC and SS-MTC output plausible replacement phrases for sensitive spans when incorporated into a two-stage text anonymization pipeline, while best preserving downstream utility (§4.3). When compared to alternative MLM infilling approaches or to DP and LLM-based approaches, our approach performs considerably better.

An ideal anonymization pipeline not only produces high-quality pseudoanonymized text, but also is lightweight, allowing practitioners to locally anonymize data with minimal compute prior to sharing it. Although the initial construction of SS-MTC requires running the detection model over the large Wikipedia corpus, which adds compute time, its generalizability across domains (§4.4) suggests that this approach has high potential for facilitating local anonymization that improves both utility and privacy (as suggested by the HIPS theory (Carrell et al., 2012)). This further motivates work on improving sensitive candidate detection, robustness to imperfect redaction models, and preserving utility under varying levels of redaction.

Finally, we note that the importance of plausible replacements depends on the downstream use case of anonymized data. While our downstream utility tasks (§4.3) offer evidence that MTC and SS-MTC outperform alternatives, differences in model performance vary across tasks and are greater for coreference resolution (Table 9) than summarization (Table 7). Plausible replacements are likely more important for tasks that particularly target entities and syntax, which are of interest in domains with private data (Gandhi et al., 2023).

7 Conclusions

We propose effective methods for replacing sensitive spans in text with realistic alternatives. Our methods enable pseudoanonymization when incorporated in a two-stage detection and replacement pipeline. By focusing on preserving privacy in text, our work aims to facilitate more responsible and ethical development of AI in domains with sensitive data. Text anonymization that preserves data quality is key to enabling transparency and accountability without compromising privacy.

Acknowledgments

The authors would like to thank the reviewers for their helpful feedback. This work was supported in part by the AI2050 Fellowship program by Schmidt Sciences and by the JHU + Amazon Initiative for Interactive AI (AI2AI).

8 Limitations

The primary limitation of our work is that we conduct our evaluations to two domains. While we expect our results to readily generalize to other domains, we cannot conclude generalizability without further evaluations. Furthermore, while our results suggest that our infilling method at least preserves the anonymization ability of the underling redaction model and likely improves privacy under the HIPS theory, we cannot guarantee that running our pipeline ensures full anonymization making data safe to release. Thus, we discourage practitioners from relying solely on our model off-the-shelf for anonymizing text.

9 Ethical Considerations

Although our work focuses on sensitive data, we run all experiments on existing public datasets, which reduces risks of our work. There is a minor risk of our methods being used in deployed contexts without in-domain evaluation. We caution that our approach should not be assumed to generalize to new data, especially the sensitive span detector model. Running our pipeline on new data without in-domain evaluation could risk leaking private information.

References

Allison Adams, Eric Aili, Daniel Aioanei, Rebecca Jonsson, Lina Mickelsson, Dagmar Mikmekova, Fred Roberts, Javier Fernandez Valencia, and Roger Wechsler. 2019. AnonyMate: A toolkit for anonymizing unstructured chat data. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 1–7, Turku, Finland. Linköping Electronic Press.

Alyaa Alfalahi, Sara Brissman, and Hercules Dalianis. 2012. Pseudonymisation of personal names and other phis in an annotated clinical swedish corpus. In *Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) Held in Conjunction with LREC*, pages 49–54.

Ahmed Musa Awon, Yun Lu, Shera Potka, and Alex Thomo. 2025. CluSanT: Differentially private and semantically coherent text sanitization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3676–3693, Albuquerque, New Mexico. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

David Carrell, Bradley Malin, John Aberdeen, Samuel Bayer, Cheryl Clark, Ben Wellner, and Lynette Hirschman. 2012. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *Journal of the American Medical Informatics Association*, 20(2):342–348.

Aipeng Chen, Jitendra Jonnagaddala, Chandini Nekkantti, and Siaw-Teng Liaw. 2019. Generation of surrogates for de-identification of electronic health records. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 70–73. IOS Press.

Sai Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. 2023. A customized text sanitization mechanism with differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5747–5758, Toronto, Canada. Association for Computational Linguistics.

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.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2024. Reducing privacy risks in online self-disclosures with language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13732–13754, Bangkok, Thailand. Association for Computational Linguistics.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 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.

Anjalie Field, Amanda Coston, Nupoor Gandhi, Alexandra Chouldechova, Emily Putnam-Hornstein, David Steier, and Yulia Tsvetkov. 2023. Examining risks of racial biases in nlp tools for child protective services. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1479–1492, New York, NY, USA. Association for Computing Machinery.

Nupoor Gandhi, Anjalie Field, and Emma Strubell. 2023. Annotating mentions alone enables efficient domain adaptation for coreference resolution. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10543–10558, Toronto, Canada. Association for Computational Linguistics.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet. *Circulation*, 101(23):e215–e220.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Fadi Hassan, Josep Domingo-Ferrer, and Jordi Soria-Comas. 2018. Anonymization of unstructured data via named-entity recognition. In *Modeling Decisions for Artificial Intelligence: 15th International Conference, MDAI 2018, Mallorca, Spain, October 15–18, 2018, Proceedings*, page 296–305, Berlin, Heidelberg. Springer-Verlag.

Lynette Hirschman and John Aberdeen. 2010. Measuring risk and information preservation: Toward new metrics for de-identification of clinical texts. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 72–75, Los Angeles, California, USA. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Zilyu Ji, Yuntian Shen, Jionghao Lin, and Kenneth R. Koedinger. 2025. Enhancing the de-identification of personally identifiable information in educational data. *Journal of Educational Data Mining*, 17(2):55–85.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Oren Kalinsky, Guy Kushilevitz, Alexander Libov, and Yoav Goldberg. 2023. Simple and effective multi-token completion from masked language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2356–2369, Dubrovnik, Croatia. Association for Computational Linguistics.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 14–19, Online. Association for Computational Linguistics.

Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. 2023. Harnessing large-language models to generate private synthetic text. *arXiv preprint arXiv:2306.01684*.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363.

Nuno Mamede, Jorge Baptista, and Francisco Dias. 2016. Automated anonymization of text documents. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 1287–1294.

Justus Mattern, Zhijing Jin, Benjamin Weggenmann, Bernhard Schoelkopf, and Mrinmaya Sachan. 2022. Differentially private language models for secure data sharing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4873, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Omri Mendels, Coby Peled, Nava Vaisman Levy, Sharon Hart, Tomer Rosenthal, Limor Lahiani, et al. 2018. Microsoft Presidio: Context aware, pluggable and customizable pii anonymization service for text and images.

Ishna Neamatullah, Margaret M Douglass, Li-Wei H Lehman, 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 Med. Inform. Decis. Mak.*, 8(1):32.

Bhanudas Suresh Panchbhai and Varsha Makarand Pathak. 2022. A systematic review of natural language processing in healthcare. *Journal of Algebraic Statistics*, 13(1):682–707.

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, Bangkok, Thailand. Association for Computational Linguistics.

Pranav Putta, Ander Steele, and Joseph W Ferrara. 2023. Differentially private conditional text generation for synthetic data production.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Krithika Ramesh, Nupoor Gandhi, Pulkit Madaan, Lisa Bauer, Charith Peris, and Anjalie Field. 2024. Evaluating differentially private synthetic data generation in high-stakes domains. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15254–15269, Miami, Florida, USA. Association for Computational Linguistics.

Bruno Ribeiro, Vitor Rolla, and Ricardo Santos. 2023. INCOGNITUS: A toolbox for automated clinical notes anonymization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 187–194, Dubrovnik, Croatia. Association for Computational Linguistics.

Patrick Ruch, Robert H. Baud, Anne-Marie Rassinoux, Pierrette Bouillon, and Gilbert Robert. 2000. Medical document anonymization with a semantic lexicon. *Proceedings of American Medical Informatics Association Symposium*, pages 729–733.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2025. Language models are advanced anonymizers. In *International Conference on Learning Representations*, volume 2025, pages 98558–98598.

Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1. *Journal of Biomedical Informatics*, 75:S4–S18.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification. *Journal of Biomedical Informatics*, 58:S20–S29.

Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the scrub system. *Proceedings of Americal Medical Informatics Association Annual Fall Symposium*, pages 333–337.

Santosh T.y.s.s., Mahmoud Aly, and Matthias Grabmair. 2024. LexAbSumm: Aspect-based summarization of legal decisions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10422–10431, Torino, Italia. ELRA and ICCL.

Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 19(5):786–791.

Arpita Vats, Zhe Liu, Peng Su, Debjyoti Paul, Yingyi Ma, Yutong Pang, Zeeshan Ahmed, and Ozlem Kalinli. 2024. Recovering from privacy-preserving masking with large language models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10771–10775.

Elena Volodina, Yousuf Ali Mohammed, Sandra Derbring, Arild Matsson, and Beata Megyesi. 2020. Towards privacy by design in learner corpora research: A case of on-the-fly pseudonymization of Swedish learner essays. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 357–369, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Oleksandr Yermilov, Vipul Raheja, and Artem Chernodub. 2023. Privacy- and utility-preserving NLP with anonymized data: A case study of pseudonymization. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 232–241, Toronto, Canada. Association for Computational Linguistics.

Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. 2021. Differential privacy for text analytics via natural text sanitization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3853–3866, Online. Association for Computational Linguistics.

Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2023. Synthetic text generation with differential privacy: A simple and practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1321–1342, Toronto, Canada. Association for Computational Linguistics.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal

artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

A Additional Experiments

We conduct the automated and downstream utility evaluation over a sweep of ϵ values for all 3 DP approaches: SanText, CusText, CluSanT in Table 10 and Table 11. CluSanT performs the best in all the experiments. While for automated evaluations (Table 10) $\epsilon = 3$ turns out to be the best choice, for downstream utility evaluation (Table 11) there is no consensus in ϵ values.

Method	ϵ	Perplexity	CI
SanText+	1	128.72	(117.38, 140.50)
	2	120.11	(110.34, 130.28)
	3	**119.48**	(109.15, 131.58)
CusText+	1	84.69	(78.95, 92.27)
	2	85.82	(78.82, 93.50)
	3	**78.40**	(74.25, 82.82)
CluSanT	1	26.55	(25.34, 28.16)
	2	26.51	(25.40, 28.01)
	3	**26.42**	(25.10, 27.97)

Table 10: Automated evaluation on language modeling task using the LexAbSumm dataset. Perplexity of the text generated by LMs trained on text pseudoanonymized by each DP approach (with different ϵ values) is evaluated against the LM trained on the Original data. Numbers inside the parentheses indicate 95% confidence interval computed through bootstrapping over the test set.

Method	ϵ	Perplexity	CI
SanText+	1	37.57	(35.77, 39.12)
	2	**37.57**	(36.05, 39.51)
	3	37.57	(35.96, 39.24)
CusText+	1	30.09	(29.05, 31.33)
	2	29.37	(28.28, 30.69)
	3	**28.30**	(27.34, 29.32)
CluSanT	1	**15.61**	(13.66, 18.19)
	2	17.78	(16.17, 19.46)
	3	16.29	(14.76, 18.79)

Table 11: Downstream Language Modeling evaluation on LexAbSumm dataset. Perplexity of the "Original" text is evaluated against each DP (with different ϵ values) approach's LM. Numbers inside the parentheses indicate 95% confidence interval computed through bootstrapping over the test set.

B Human Evaluation

We use 2 kinds of instructions for each task, one introducing the task with examples and the second with each question reminding the annotators of the core task. We list the introductory instructions used in Task 1, and Task 2 in Figure 1 and Figure 2, respectively. Sample questions for Task 1 and Task 2 are in Figure 3 and Figure 4

Our goal in this project is to evaluate tools for anonymizing text. This section of the survey wants you to choose the **Best** and the **Worst** among 4 different anonymized versions of a sentence taken from a legal document, where highlighted phrases in the original text have been replaced by alternatives. Base your choice on:

- overall grammatical correctness,
- whether the replacement matches the function of the original phrase (a name is replaced with a different name, a place with a different place, etc.)
- whether the overall meaning of the sentence remains the same

Here are some examples with reasons why one version is better than the other.

Example 1

Original
*The Court heard addresses by **Mr. Corell** for the Government and by **Mr. Gaukur Jörundsson** for the Commission, as well as their replies to the questions put by the Court and several judges.*

Anonymized Options
A: *The Court heard addresses by **the, Secretary** for the Government and by **the Deputy Secretary** for the Commission, as well as their replies to the questions put by the Court and several judges.*
B: *The Court heard addresses by **the applicant, acting** for the Government and by **the Minister for the,,, and and and** for the Commission, as well as their replies to the questions put by the Court and several judges.*
C: *The Court heard addresses by **lawyers** for the Government and by **lawyers** for the Commission, as well as their replies to the questions put by the Court and several judges.*
D: *The Court heard addresses by **1848** for the Government and by **1848** for the Commission, as well as their replies to the questions put by the Court and several judges.*
Solution
C sounds the most coherent and keeps the sentence's intent still the same. B sounds the least coherent with repeated subsequent commas and extra white spaces.

Example 2

Original
*An action, the object of which was to settle the terms of the expropriation (see paragraph 15 below), was commenced on **28 February 1980** before the **Real Estate Court** at the **Falun District Court (tingsrätten)**.*

Anonymized Options
A: *An action, the object of which was to settle the terms of the expropriation (see paragraph 15 below), was commenced on **15 February,** before the **Real Court Court** at the **end of Court**.*
B: *An action, the object of which was to settle the terms of the expropriation (see paragraph 15 below), was commenced on **15 February,** before the **Real Court Court** at the **end of of the (((, below)**.*
C: *An action, the object of which was to settle the terms of the expropriation (see paragraph 15 below), was commenced on **29 January** before the **public hearing** at the **Municipal**.*
D: *An action, the object of which was to settle the terms of the expropriation (see paragraph 15 below), was commenced on **Broadway** before the **1867** at the **Netherlands**.*
Solution
C is the least grammatically wrong answer, so it comes out at top. B has serious issues with the grasp of grammar, as it repeats words and chains multiple punctuation marks one after the other. So, B is the last ranked. Others are better than B as their replacements are valid independent phrases.

Figure 1: Introductory instruction for Task 1 of Human Evaluation

Our goal in this project is to evaluate tools for anonymizing text. This section of the survey wants you to evaluate 4 different anonymized versions of a sentence taken from a legal document, where highlighted phrases in the original text should be replaced by alternatives. **We only care about the highlighted parts** of the sentences, parts that aren't highlighted shouldn't be considered while answering. Choose whether each version is anonymized or not, based on whether the replacement ends up revealing the original entity.

Here are some examples with reasons why the version is anonymized or not.

Example 1

Original
*for the Commission **Mr. Gaukur Jörundsson**, Delegate. The Court heard addresses by **Mr. Corell** for the Government and by **Mr. Gaukur Jörundsson** for the Commission, as well as their replies to the questions put by the Court and several judges.*

Anonymized Option
*for the Commission **Adviser**, Delegate. The Court heard addresses by **the applicant**, acting for the Government and by **the Minister for the,,, and and and** for the Commission, as well as their replies to the questions put by the Court and several judges.*
Solution
Yes, all the sensitive information is replaced with non-identifying information.

Anonymized Option
*for the Commission **Mr. Gaukur Jörundsson**, Delegate. The Court heard addresses by **the applicant**, acting for the Government and by **the Minister for the,,, and and and** for the Commission, as well as their replies to the questions put by the Court and several judges.*
Solution
No, as the identity of the the Commission's delegate hasn't been anonymized.

Anonymized Option
*for the Commission **Mr. Jörundsson**, Delegate. The Court heard addresses by **the applicant**, acting for the Government and by **the Minister for the,,, and and and** for the Commission, as well as their replies to the questions put by the Court and several judges.*
Solution
No, as the identity of the the Commission's delegate hasn't been anonymized.

Anonymized Option
*for the Commission **of)) law ,, etc. blah**, Delegate. The Court heard addresses by **the applicant**, acting for the Government and by **the Minister for the,,, and and and** for the Commission, as well as their replies to the questions put by the Court and several judges.*
Solution
Yes, all the sensitive information is replaced with non-identifying information as we can't identify the original names.

Anonymized Option
*for the Commission **28th September 2012**, Delegate. The Court heard addresses by **the applicant**, acting for the Government and by **the Minister for the,,, and and and** for the Commission, as well as their replies to the questions put by the Court and several judges.*
Solution
Yes, all the sensitive information is replaced with non-identifying information as we can't identify the original names.

Example 2

Original
*The **Istanbul State Security Court** held **twenty-eight more** hearings before delivering its final judgment.*

Anonymized Option
*The **State Security Court** held **four further** hearings before delivering its final judgment.*
Solution
Yes, all the sensitive information is replaced with non-identifying information, as it doesn't reveal the name of the court.

Figure 2: Introductory instruction for Task 2 of Human Evaluation

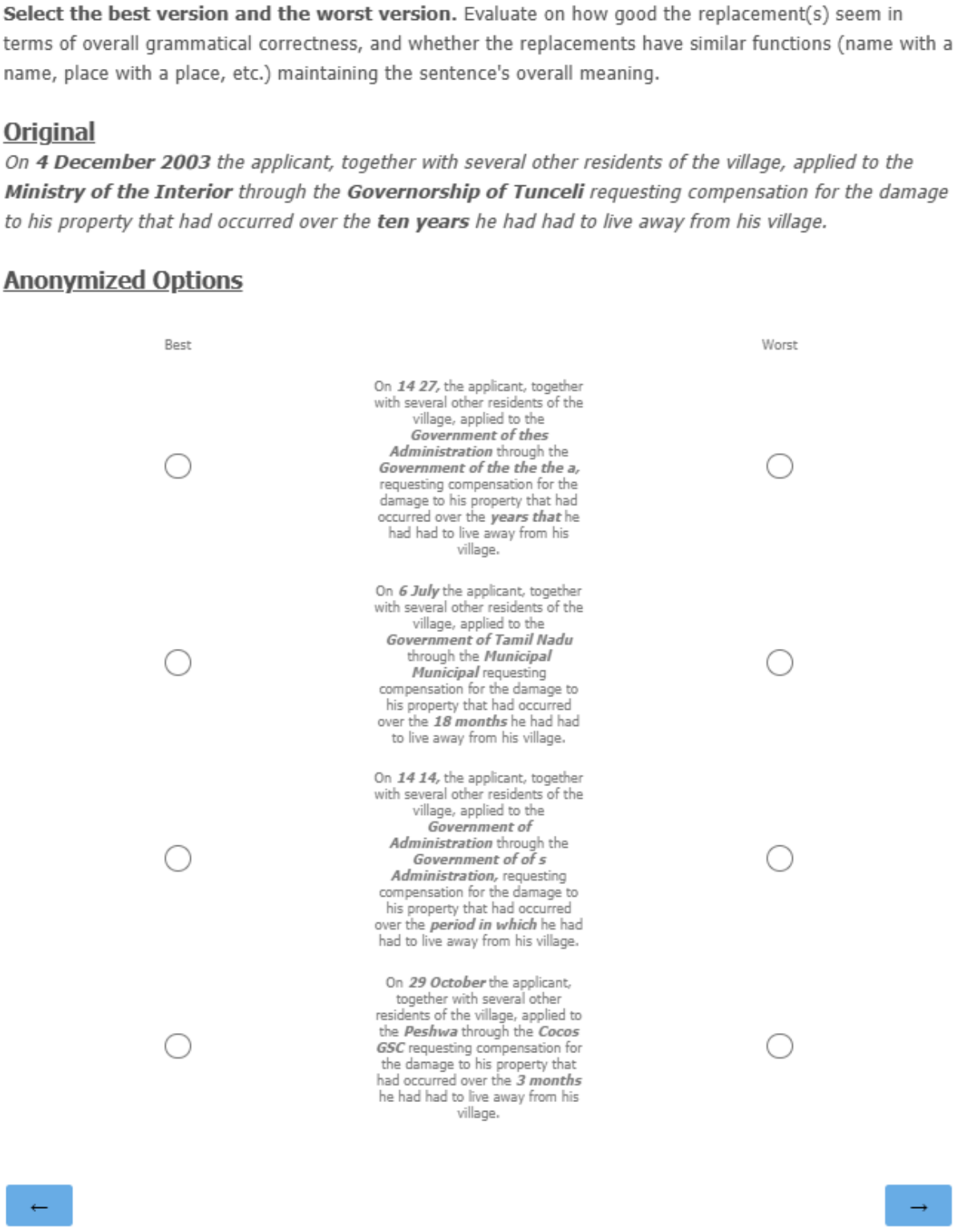

Figure 3: Survey question from bucket 1 of Task 1 of Human Evaluation

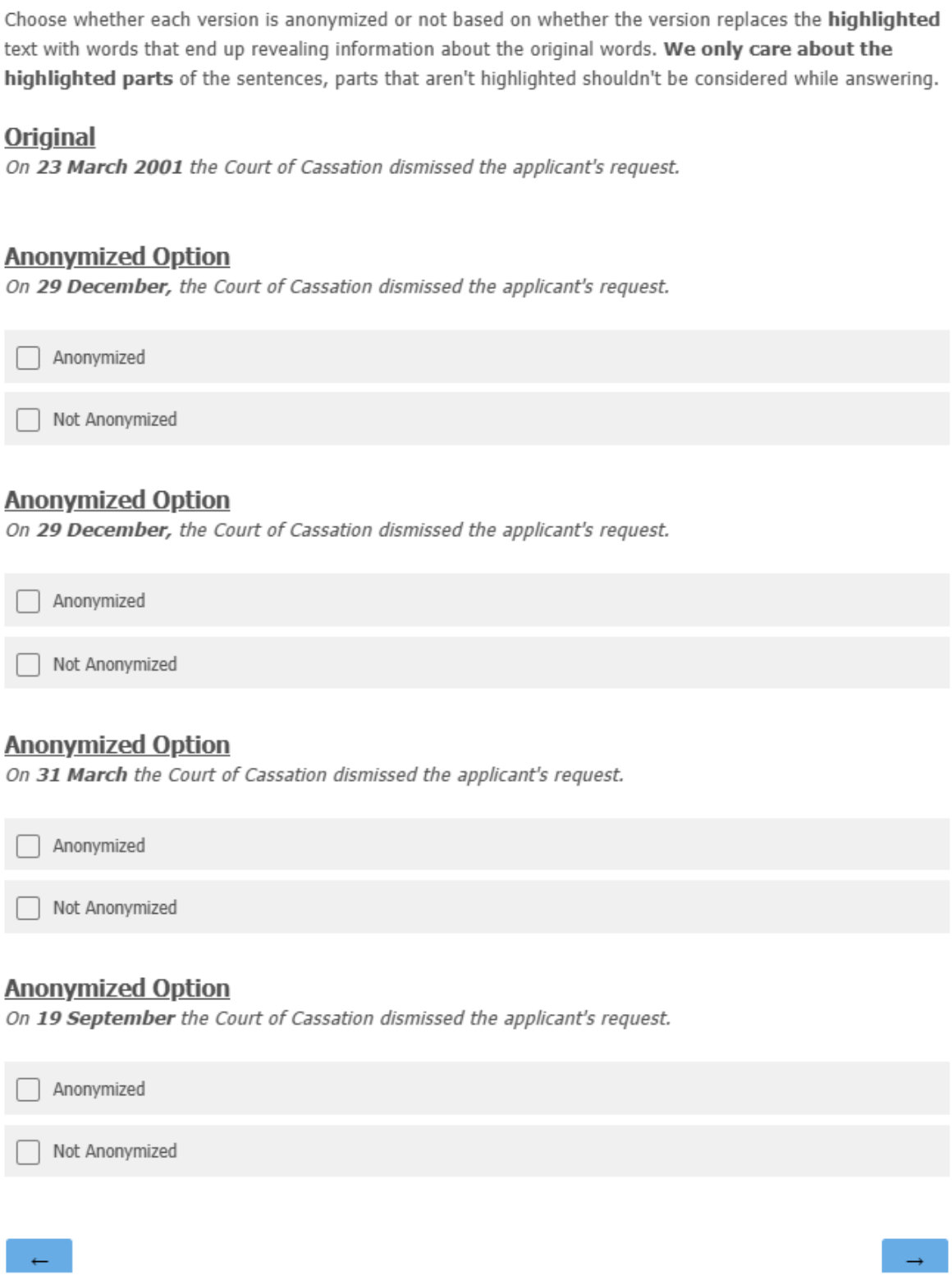

Figure 4: Survey question from bucket 1 of Task 2 of Human Evaluation