Granularity is crucial when applying differential privacy to text: An investigation for neural machine translation

Doan Nam Long Vu[1] and **Timour Igamberdiev**[1] and **Ivan Habernal**[2]
Trustworthy Human Language Technologies
[1] Department of Computer Science, Technical University of Darmstadt
[2] Research Center Trustworthy Data Science and Security of the University Alliance Ruhr, Faculty of Computer Science, Ruhr University Bochum
timour.igamberdiev@tu-darmstadt.de
www.trusthlt.org

Abstract

Applying differential privacy (DP) by means of the DP-SGD algorithm to protect individual data points during training is becoming increasingly popular in NLP. However, the choice of granularity at which DP is applied is often neglected. For example, neural machine translation (NMT) typically operates on the sentence-level granularity. From the perspective of DP, this setup assumes that each sentence belongs to a single person and any two sentences in the training dataset are independent. This assumption is however violated in many real-world NMT datasets, e.g., those including dialogues. For proper application of DP we thus must shift from sentences to entire documents. In this paper, we investigate NMT at both the sentence and document levels, analyzing the privacy/utility trade-off for both scenarios, and evaluating the risks of not using the appropriate privacy granularity in terms of leaking personally identifiable information (PII). Our findings indicate that the document-level NMT system is more resistant to membership inference attacks, emphasizing the significance of using the appropriate granularity when working with DP.[1]

```
{
...
"de": "Kunde: Immo Hande-Hornig",
"en": "Customer: Immo Hande-Hornig",
...
"de": Agent: ... Ich bin Immo Hande-Hornig.
"en": Agent: ... you are through to Immo Hande-Hornig.
...
}
```

Figure 1: Examples of sentences that are *not independent* within a document. The independence is violated via the "Immo Hande-Hornig" sequence, breaking the DP guarantee of protecting each sentence during the training process.

1 Introduction

With increasing concerns about the privacy of individuals and data leakage from NLP systems (Carlini et al., 2021), a method that has gained popularity in privacy-preserving NLP is Differential Privacy (DP) (Hu et al., 2024). However, the exact manner in which DP is applied to a textual dataset has numerous pitfalls. The *unit of privacy* is one among them, i.e. the granularity at which we assume an individual 'data point' (e.g. sentences, documents, and so forth) (Ponomareva et al., 2022; Igamberdiev and Habernal, 2023), with an assumption of *independence* among data points (Dwork and Roth, 2013).

One particular task that has recently raised many privacy concerns is neural machine translation (NMT). Applying DP at the sentence level for NMT may break the independence assumption if more than one sentence is associated with a single individual (Brown et al., 2022), as depicted in Figure 1. In such cases, scaling up the unit of privacy to the document level by grouping related sentences overcomes the violated privacy protection which 'pretends' all sentences are independent, i.e. the status quo.

The main objective of this paper is to compare the use of DP for NMT systems and datasets at the sentence and document levels, focusing on the level of privacy protection offered. First, we investigate the trade-off between privacy and utility for different levels of granularity (sentence vs. document) during the training process of an NMT system. Specifically, we examine how performance is affected when applying DP with varying levels of privacy guarantees and granularity. Secondly, we aim to evaluate the risks of not using a proper privacy granularity during the training process of an NMT system through data extraction attacks on these systems.

Our contributions are as follows. (1) We propose a novel approach to apply differential privacy to NMT systems at the document level, uti-

[1]Our code is available at https://github.com/trusthlt/granularity-is-crucial-dp.

Findings of the Association for Computational Linguistics: EACL 2024, pages 507–527
November 12-16, 2024 ©2024 Association for Computational Linguistics

lizing the DP-NMT framework (Igamberdiev and Habernal, 2023) and the mLongT5 model (Uthus et al., 2023). The evaluation results show that the document-level NMT system is extremely sensitive to the privacy budget (ε), which can significantly affect performance and lead to a drop in utility. We therefore suggest training the document-level NMT system on a larger, non-sensitive dataset, such as WMT22 (Kocmi et al., 2022), to achieve a better trade-off between privacy and utility on the downstream dataset with DP. (2) We apply the loss-based membership inference attack (MIA) (Yeom et al., 2018) to detect private information in NMT systems at both the sentence and document levels. Based on this MIA, we create an evaluation schema for personally identifiable information (PII) to estimate the percentage of potential information leakage. Our results show that the document-level NMT system is more robust against the loss-based MIA than the sentence-level NMT system, demonstrating the importance of using the proper granularity when working with DP.

2 Background

2.1 Differential Privacy

We refer to Appendix A for a detailed explanation of Differential Privacy and DP-SGD.

A key aspect of applying DP to text is that we cannot simply utilize group privacy (Dwork and Roth, 2013) to achieve document-level privacy from sentence-level privacy. Since DP-SGD leverages the Approximate DP definition (Dwork and Roth, 2013), the δ value, i.e. the probability of privacy leakage occurring, is not 0 when applying group privacy. If the number of data points k in which two neighboring datasets differ is large enough, δ will exceed 1 due to scaling with a factor of $ke^{(k-1)\varepsilon}$. Moreover, using relaxed DP definition such as Renyi DP (Mironov, 2017) to avoid including the δ value will scale ε to a very large value that is practically unmanageable, also resulting in a very weak privacy guarantee.

2.2 Overview of membership inference attacks (MIA)

Since datasets used to train neural models often contain confidential user information, they may be vulnerable to privacy risks (Hu et al., 2022). The trained models are often over-paramet erized, meaning they can memorize information about their training dataset (Mireshghallah et al., 2022). They exhibit a different behavior on training data compared to test data, with model parameters storing information about specific training data unit. Membership inference attacks (MIAs) aim to predict whether specific examples are members of the training dataset (Hu et al., 2022). In general, an MIA is actually a binary classifier, which is designed to distinguish a target model's behavior of its training members from the non-members.

The first MIA was proposed by Shokri et al. (2017), utilizing shadow datasets that have similar distribution to the original training data. Multiple shadow models are trained on these datasets, which are meant to mimic the behavior of the target model. The output predictions of these models are then used as input to a final binary classification model. This model detects whether a given data point belongs to the target model or not.

3 Related Work

3.1 Previous work on DP+NLP

Several works attempted to pre-train language models with DP-SGD (Anil et al., 2022; Yin and Habernal, 2022; Ponomareva et al., 2022). With a significant computational burden of pre-training with DP-SGD, finetuning pre-trained language models using DP-SGD has seen increased research over the past few years (Senge et al., 2022; Li et al., 2022). The main objective is to utilize a pre-trained checkpoint of a model that was created by using a publicly available corpus of data, such as Wikipedia or C4 (Raffel et al., 2020) and then fine-tune it on a private downstream dataset with DP-SGD. Although fine-tuning is more efficient than pre-training, it still requires a large amount of data on text generation tasks to achieve good performance, e.g. language modeling (Li et al., 2022) or NMT (Igamberdiev et al., 2024). The aforementioned related works only concentrate on the protection of the gradient at the *sentence level*. In contrast, the current work is concerned with the protection of the gradient at the *document level*.

3.2 Previous work on document-level NMT

Sentence-level NMT is the most common method of machine translation because it is simpler to train and evaluate with existing large datasets and evaluation metrics (Post and Junczys-Dowmunt, 2023). The primary limitation is due to the NMT model's memory consumption for sequence length, resulting in a larger memory footprint when increased.

Secondly, popular datasets, such as WMT (Fernandes et al., 2021), exist only for sentence-level machine translation, even though they were originally created as documents. Nonetheless, this approach has its limitations, as demonstrated in the current study, which is the related privacy issue with DP-SGD at the sentence level. Moreover, from the perspective of machine translation, a sentence-level MT model is not suitable for translating lengthy documents without taking their context into account (Wicks and Post, 2023; Wu et al., 2023). Recent related works on document-level machine translation can be separated into two categories: *Encoder-Decoder Models* and *Decoder Only Models*.

Encoder-Decoder Models Most of the recent works focus on the standard Transformer model (Wu et al., 2023; Zhuocheng et al., 2023a). Typically, they concatenate multiple sentences to form a document with the length up to 512 or maximum 1024 tokens for training. However, the naive approach generally suffers from a **length bias problem**, which causes significant degradation in translation quality when decoding documents that are much shorter or longer than the maximum sequence length during training (Zhuocheng et al., 2023a).

Decoder Only Models Unlike the vast majority of the training/fine-tuning paradigm, recent works (Hendy et al., 2023; Karpinska and Iyyer, 2023) suggest that Generative Pre-Training (GPT) models (Radford et al., 2018) are able to achieve very competitive translation quality on document-level translation. As such, they use a few-shot prompting technique to translate a document, employing ChatGPT[2]. The prompt displays examples of each translated sentence pair first and instructs the model to consider the context when translating, as in a document.

4 Methods

4.1 Document-level machine translation

We employ the DP-NMT framework developed by Igamberdiev et al. (2024) for privacy-preserving NMT with DP-SGD. The framework is built on top of Flax (Heek et al., 2023) and JAX (Bradbury et al., 2018) for rapid DP-SGD training. By default, the framework supports models such as mBART (Liu et al., 2020) or mT5 (Xue et al., 2021) out of the box. However, these multilingual seq2seq models are not suitable for our task, as they are pre-trained on shorter sentences.

mLongT5 model mLongT5 is a multilingual seq2seq model based on LongT5 (Guo et al., 2022), which is a seq2seq model that uses T5 (Raffel et al., 2020) as its foundation, with a *Transient Global (TGlobal) Attention* mechanism. This attention mechanism is well-suited for long text tasks in terms of memory efficiency, and mLongT5 leverages the mC4 dataset (Xue et al., 2021) with 4096 token long input sequences for pre-training. As mLongT5 checkpoints are designed for long text tasks and available for Flax and JAX, we incorporate mLongT5 into the DP-NMT framework.

4.2 Loss-based MIA

Previous work attacking an NMT system used the entire WMT dataset to create a sophisticated shadow MIA attack (Hisamoto et al., 2020). However, in our work, the datasets on which we conduct experiments (see Section 5) have less than 20,000 data points, which is significantly fewer than the millions of data points in WMT. This makes shadow MIA less effective. Also, in terms of computational complexity for the shadow MIA, an attacker must train hundreds of shadow models to achieve good performance (Yeom et al., 2018). Loss-based metrics (Yeom et al., 2018) are less computationally intensive to perform. Intuitively, if the loss of a data point is smaller than the target model's expected training loss, the record is classified as a member; otherwise as a non-member. The target model is trained by minimizing the prediction loss of its training members. Therefore, the prediction loss of a training record should be smaller than that of a test record. The attack $\mathsf{Exp}^{\mathsf{M}}_{\mathsf{loss}}$ is defined as follows:

$$\mathsf{Exp}^{\mathsf{M}}_{\mathsf{loss}} = \mathbb{1}(\ell(\theta(\mathbf{r}|\mathbf{s}); \mathbf{r}) \leq \tau), \qquad (1)$$

where θ is the model, $\mathbf{r}$ is target output, $\mathbf{s}$ is the input, ℓ is the loss function (typically cross-entropy loss), τ is a threshold (average training loss) and $\mathbb{1}$ is a classifier function, which takes event A and returns 1 if the event A occurs, 0 otherwise.

Loss-based MIA highlights that **overfitting** of target ML models is the primary factor contributing to the success of MIAs. This attack *strongly exploits* the different behaviors of target ML models on their training versus test data. The attacker is

[2] https://openai.com/blog/chatgpt

assumed to have knowledge of the data points, but it is uncertain whether they were used in training. In fact, we are assuming a very powerful adversary that already has knowledge of the data, which is why this is a strong white-box attack for investigating data leakage. The original evaluation metric scheme for the loss-based MIA is the attack advantage or privacy leakage resulting from differences between the false positive rate (FPR) and true positive rate (TPR) of the attack: $\mathsf{Adv}^{\mathsf{M}} = \mathrm{TPR} - \mathrm{FPR}$

4.3 PII exposure

Considering *private information* as named entities that a model might overfit to during the training process, we present a *evaluation scheme* to test the effectiveness of the MIA with respect to PII. The method works as follows: We carry out the loss-based MIA, obtaining extracted training records from Eqn. 1. We then select the **true positive predictions** and calculate the number of PII that are present. For sentence-level privacy, due to correlation of sentences within a document, more PII leakage would be expected than with document-level privacy. This aims to see how well the model optimizes against entities. We use the default pipeline setting of `Presidio`[3] which consists of RegEx, Spacy NER and BERT contextual awareness to extract a set of PII from given sentences.[4]

5 Experiments

5.1 Datasets

We aim to find a suitable dataset which mimics the real private environment of processing sensitive information, but is publicly available, both for reproducibility and ethical reasons. In addition, a dataset must be appropriate for both sentence-level and document-level machine translation. Therefore, we select three datasets for our investigations, BSD, MAIA and Europarl, described below.

BSD The Business Scene Dialogue corpus (BSD) (Rikters et al., 2019) is a collection of fictional business conversations in various scenarios, with parallel data for Japanese and English. For our experiments, we combined the original corpus, which consists of two translation direction into a single Japanese → English (JA-EN) language pair.

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

[4] We note that `Presidio` is not as thorough as human annotators, and may miss some PII data or return false positives. However, it is still a good approximation of the number of PII.

Dataset	Level	# Train	# Val.	# Test
BSD	Sentence	20,000	2,120	2,051
	Document	670	69	69
MAIA	Sentence	13,380	2,488	2,109
	Document	355	71	70
Europarl	Sentence	1,454,229	181,774	181,764
	Document	143,706	19,786	19,967

Table 1: Number of training examples for both datasets in sentence-level and document-level

MAIA The Multilingual Artificial Intelligence Agent Assistant (MAIA) corpus consists of genuine bilingual (German-English) customer support conversations from the Unbabel database (Farinha et al., 2022). To make the conversations publicly available, the data was first anonymized using the Unbabel proprietary anonymization tool and then manually verified.

Europarl The Europarl (Koehn, 2005) dataset is a widely used parallel corpus in the field of machine translation. Extracted from the proceedings of the European Parliament, the dataset includes versions in 21 European languages. For this work, we use the Europarl V10 version from WMT22 (Kocmi et al., 2022) for German-English bilingual text.

5.2 Data preparation

Since those datasets are dialogue/speech session datasets, we need to concatenate the utterances within a dialogue/speech session into a single document for document-level machine translation. First, we concatenate the speaker's name and the utterance into a single sentence. Namely, `<SPEAKER>: <UTTERANCE>`. Then we concatenate all the utterances within a dialogue/speech session into a single document. This process results in a smaller number of training examples than the original sentence-level training examples. Table 1 shows the number of training examples for each datasets at the sentence level and document level. We refer to Figures 6, 7 and Figure 8 in Appendix B for examples of each dataset preparation.

5.3 Experimental setup

The training experiment directly on Huggingface's mLongT5 checkpoint[5] is denoted as θ^{sen} at the sentence level, and as θ^{doc} at the document level. We denote training data at the sentence level

[5] https://huggingface.co/agemagician/mlong-t5-tglobal-base

as $\mathcal{D}^{\mathsf{sen}}_{\mathrm{train}}$ and at the document level as $\mathcal{D}^{\mathsf{doc}}_{\mathrm{train}}$, similarly for validation ($\mathcal{D}^{\mathsf{sen}}_{\mathrm{val}}$, $\mathcal{D}^{\mathsf{doc}}_{\mathrm{val}}$) and test data ($\mathcal{D}^{\mathsf{sen}}_{\mathrm{test}}$, $\mathcal{D}^{\mathsf{doc}}_{\mathrm{test}}$). We refer to Table 3 in Appendix for the notation and its description used in this work.

Additional pre-training with WMT22 The number of document-level training examples is much smaller than at the sentence level (See Table 1). Thus, the model may underfit the training data during private training, resulting in poorer performance than normal training at the document level and private training at the sentence level. To improve the document-level model performance during private training, we fine-tuned mLongT5 checkpoint without DP-SGD on the WMT22 dataset first before fine-tuning on downstream datasets at the document level. After fine-tuning on the document-level WMT22 dataset, we fine-tune the model on the BSD and MAIA datasets at the document level for both normal and private training. We also use the pre-trained checkpoint on WMT22 with Europarl for comparison. The model training experiments based on the document-level WMT22 dataset are denoted as $\theta^{\mathrm{augdoc}}_{\mathrm{zero-shot}}$ and with downstream data as θ^{augdoc}.

5.4 Hyperparameters

The primary distinction between two level models in terms of hyperparameters is the maximum sequence length. For sentence-level training, it is set to 64-128, whereas for document-level training, it is set to 1200-1500. We refer to the details of our hyperparameters search in Appendix C.

For additional training with WMT22, we use the same hyperparameters as document-level settings. To prepare the document-level WMT22 dataset, we concatenate the multiple sentences into a single document, as long as they reach 1200 tokens for Japanese to English and 1600 tokens for German to English. Those documents are aligned with the original sentence-level training examples. We refer to Table 7 in Appendix C.3 for the number of training examples in our experiment with WMT22.

Privacy Hyperparameters We compare ε values[6] of $\infty, 990, 90, 10$ and 1 for training on MAIA, then $\infty, 400, 40, 10$ and 1 for training on BSD and $\infty, 3130, 313, 10$ and 1 for training on Europarl. Those values are applied to both sentence-level and document-level training. We refer to the details of our privacy guarantee in Appendix H.

Given that sentences must be concatenated to form a document (see Table 1), the document-level datasets are necessarily smaller than the original ones. Consequently, the σ noise introduced during document-level training with the DP-SGD algorithm is increased and a higher sampling rate is employed, in order to match the exact same ε value among the two configurations. Apart from this, the DP-SGD hyperparameters, such as the gradient clip value C, are identical for both settings. Furthermore, we use more epochs to train document-level model in private setting to obtain decent performance (see Appendix C).

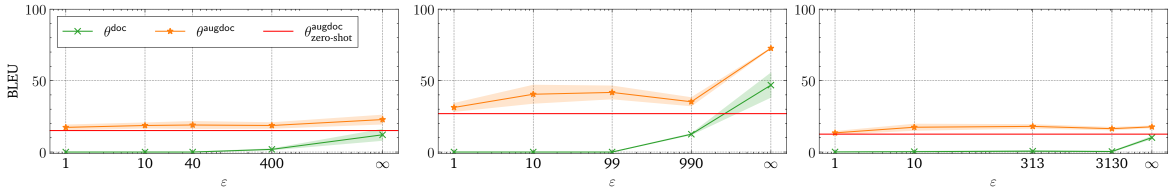
5.5 Evaluation

Performance We report BLEU (Papineni et al., 2002) for n-gram matching evaluation and BERTScore (Zhang et al., 2020) for semantic similarity evaluation. We refer to Appendix D for the details of our modification to BERTScore for evaluation on long texts.

MIA It is difficult to know whether a sentence belongs to the training data of a document-level model or a sentence-level model. For each dataset at sentence level, we consider the validation set and the test set as *non-members* and the training set as *members*. However, the training set is huge compared to the validation set and the test set; it is recommended to balance the dataset for MIA evaluation to avoid the bias of the attacker (Jayaraman and Evans, 2019). We sample the total number of members from the training set to be equal to the total number of examples in the validation set and the test set for our experiments, similar to Yeom et al. (2018); Jayaraman and Evans (2019). Formally, let α be the total number of sentences in the validation set $\mathcal{D}^{\mathsf{sen}}_{\mathrm{val}}$ and the test set $\mathcal{D}^{\mathsf{sen}}_{\mathrm{test}}$. By leveraging Sampling Without Replacement to avoid duplicating instances, we have a sampled set of sentences $\mathcal{D}^{\mathsf{sen}}_{\mathrm{sampled}}$ from a sentence-level training set $\mathcal{D}^{\mathsf{sen}}_{\mathrm{train}}$: $(s_i, r_i), \ldots, (s_\alpha, r_\alpha) \sim \mathcal{D}^{\mathsf{sen}}_{\mathrm{train}}$, where s_i is the source and r_i is the corresponding target sentence for $i \in \{1, \ldots, \alpha\}$ and $|\mathcal{D}^{\mathsf{sampled}}_{\mathrm{val}}| = |\mathcal{D}^{\mathsf{sen}}_{\mathrm{val}}| + |\mathcal{D}^{\mathsf{sen}}_{\mathrm{test}}|$ (See Table 6a in Appendix E for the exact number).

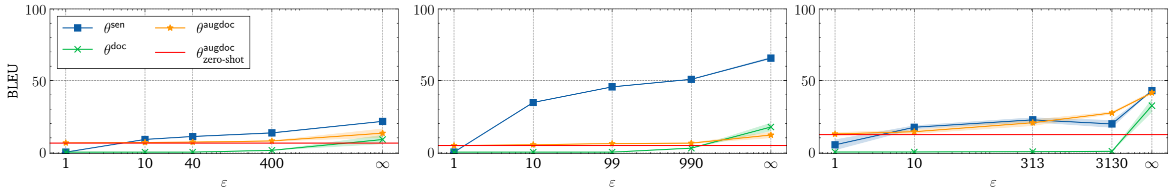
PII In NMT, the cross-lingual PII detection might be not comparable between languages of input s

[6] The maximum number of utterances in a dialogue/speech session within each dataset (99 for MAIA and 40 for BSD, 313 for Europarl) is multiplied by the ε values of 1 and 10. Finally, the resulting values are ε equals 990, 90 for MAIA, then 400, 40 for BSD and 3130, 313 for Europarl.



(a) BSD (b) MAIA (c) Europarl

Figure 2: BLEU scores on $\mathcal{D}^{\text{doc}}_{\text{test}}$ for the three document-level model fine-tuning configurations. Lower ε corresponds to better privacy.



(a) BSD (b) MAIA (c) Europarl

Figure 3: BLEU scores on $\mathcal{D}^{\text{sen}}_{\text{test}}$ for all four model fine-tuning configurations. Lower ε corresponds to better privacy.

and target r. Hence, we only report the PII leakage estimation on the target language based on the reference r (See Table 6b in Appendix for the number of PII in $\mathcal{D}^{\text{sen}}_{\text{sampled}}$ of each dataset). We use the percentage of PII leakage estimation as the metric, since all PIIs are extracted from the sampled training data. Namely, this is the ratio of the number of detected PII data to the total number of PII data in the sampled training data:

$$\text{PII}_{\%\ \text{leakage}}(\mathbf{r}) = \frac{\text{PII of TP } \mathbf{r} \in \mathcal{D}^{\text{sen}}_{\text{sampled}}}{\text{Total PII of } \mathbf{r} \in \mathcal{D}^{\text{sen}}_{\text{sampled}}} \tag{2}$$

We suspect a private training model to have a lower PII leakage percentage than 50%.

6 Results

6.1 Privacy/utility trade-off

We present the BLEU score resutls below, we refer to Appendix F.2 for BERTScore results.

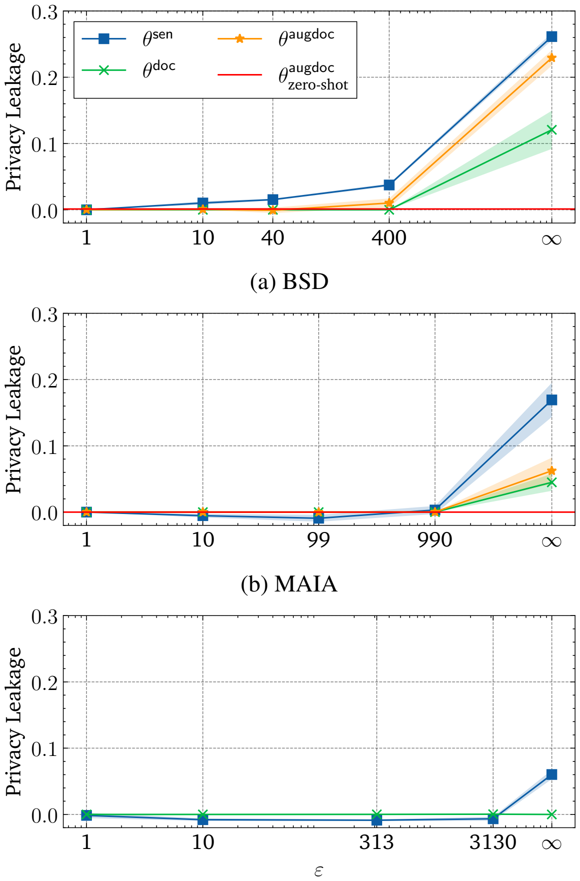
Evaluation on $\mathcal{D}^{\text{doc}}_{\text{test}}$ Figure 2 shows the BLEU score of the two approaches on $\mathcal{D}^{\text{doc}}_{\text{test}}$. As expected, we can observe the deterioration of the BLEU score as the value of ε decreases on both datasets. **The additional training data from WMT22 is beneficial for the translation quality on both datasets**. Without pre-training on WMT22, the BLEU score of θ^{doc} is significantly lower than θ^{augdoc}, with in a significant drop in translation quality. Moreover, the BLEU score of $\theta^{\text{augdoc}}_{\text{zero-shot}}$, which is fine-tuned on WMT22, is even higher than the fine-tuned θ^{doc} at $\varepsilon = \infty$ on the BSD dataset. The results of θ^{augdoc} are consistently better than θ^{doc} across all values of ε. Overall, these results indicate that privately fine-tuning on the target task is slightly beneficial for the translation quality, with respect to the domain translation quality at the document level.

Evaluation of $\mathcal{D}^{\text{sen}}_{\text{test}}$ Figure 3 shows the BLEU score of the three approaches on $\mathcal{D}^{\text{sen}}_{\text{test}}$. On MAIA, the BLEU score of θ^{sen} is superior to θ^{doc} and θ^{augdoc} at any chosen value of ε, except for $\varepsilon = 1$. Even at $\varepsilon = 10$, the BLEU score of θ^{sen} is very high at 35 BLEU score, while the BLEU score of all document-level models are low.

On BSD, the performance gap between θ^{sen}, θ^{doc} and θ^{augdoc} is less significant, possibly due to the distantly related language pair. The difference in BLEU score between θ^{sen} and θ^{augdoc} is about 10 for $\varepsilon > 10$. Pre-training on WMT22 is beneficial for the translation quality at the sentence level on the BSD dataset, since the BLEU score of $\theta^{\text{augdoc}}_{\text{zero-shot}}$ is already higher than θ^{doc} at $\varepsilon = \infty$.

On Europarl, we also observe that the results of θ^{sen}, θ^{doc} and θ^{augdoc} at $\varepsilon = \infty$ are very close. Similar to the previous results on $\mathcal{D}^{\text{doc}}_{\text{test}}$, we could not achieve any sort of performance level when training θ^{doc} with DP-SGD.

We refer to Table 8, Table 9 and Table 10 in Appendix G for specific translation examples of each dataset.



(a) BSD

(b) MAIA

(c) Europarl

Figure 4: Privacy leakage using the loss-based MIA on $\mathcal{D}^{\text{sen}}_{\text{sampled}}$, $\mathcal{D}^{\text{sen}}_{\text{val}}$ and $\mathcal{D}^{\text{sen}}_{\text{test}}$, for all four model fine-tuning configurations.

6.2 Privacy risk evaluation

As in our assumption, we consider the adversary knows the average loss of the model on the training data, more particularly, the average loss of θ^{sen} on $\mathcal{D}^{\text{sen}}_{\text{train}}$ at $\varepsilon = \infty$ before performing any attack evaluation. Figure 4 shows the privacy leakage for both MAIA and BSD using loss-based MIA.

MAIA On MAIA, **the privacy leakage via loss-based MIA on document-level training is lower than sentence-level training**. Empirically, the leakage on θ^{doc} and θ^{augdoc} is 50% lower than θ^{sen} at $\varepsilon = \infty$. Furthermore, the value continues to decrease and eventually drops below zero as the variable ε decreases as FPR is higher than TPR, results in ineffective MIA. The same applies to the other models, θ^{doc}, θ^{augdoc}, we observe no privacy leakage after applying DP due to the difficulty in optimizing longer sequences plus the noise added to the gradients according to the privacy budget.

BSD On BSD, **the attacker has a higher advantage at any level of** ε. This implies that the training data has distinct characteristics that make it easier for the adversary to infer the membership of the training data vs. test data compared to MAIA. At $\varepsilon = \infty$, the privacy leakage of θ^{augdoc} is only 0.03 points behind θ^{sen}. The privacy leakage on θ^{doc} **is also lower than** θ^{sen}; however, θ^{augdoc} still has a small leakage at $\varepsilon = 400$ and the leakage of θ^{augdoc} converges to near zero at $\varepsilon = 10$.

Europarl On Europarl,[7] we find that the attacker has only a small advantage in normal training with θ^{sen}, which may be due to the similar optimization for the domain of large training data, while θ^{doc} shows no leakage. Using DP-SGD, all models show no leakage as well.

6.3 PII disclosure

Despite DP mitigating the privacy leakage to a large extent, the true positive prediction from MIA after applying DP still plays a significant role in the privacy risk evaluation. This helps us determine **the extent to which the model could potentially reveal PII** (issue of overfitting).

MAIA The PII leakage percentage of θ^{doc} and θ^{augdoc} is 0 after applying differential privacy. At $\varepsilon = \infty$, the PII leakage percentage of those approaches is below 0.25, while θ^{sen} is approximately 0.80 at $\varepsilon = \infty$ and down to 0.40 at $\varepsilon = 10$.

BSD On BSD, surprisingly, the PII leakage percentage of θ^{augdoc} at $\varepsilon = \infty$ is slightly higher than θ^{sen}. Both are around 0.75, while θ^{doc} is approximately 0.4. The most interesting observation is that the PII leakage percentage of θ^{augdoc} deteriorates faster as ε increases compared to θ^{sen}, though it is higher at $\varepsilon = \infty$. This is a significant sign that the θ^{doc} and θ^{augdoc} is more effective in reducing privacy risks than θ^{sen}.

Europarl The results of PII leakage percentage on Europarl are similar to BSD and MAIA. θ^{doc} is 0 after training with DP-SGD, while θ^{sen} leaks progressively less at lower ε values. At $\varepsilon = 1$, instead of being 0 as on MAIA and BSD, there is a small amount of PII leakage on Europarl with θ^{sen}. However, the result is insignificant for the 50% threshold of a private training model.

[7] We do not conduct the privacy risk and PII disclosure experiments with θ^{augdoc}, since Europarl is part of the WMT dataset.

7 Discussion

7.1 Trade-off for privacy granularity

Overall on $\mathcal{D}^{\text{doc}}_{\text{test}}$, the BLEU scores of θ^{augdoc} greatly vary across different values of ε in the private training setup on both datasets. This results in higher mean and standard deviation of BLEU scores for θ^{augdoc} at the lower values of ε, compared to higher values of ε. This may be due to the private training process being unstable when optimizing the objective function with respect to long sequences, which is the case of θ^{doc} in normal training.

For the results on $\mathcal{D}^{\text{sen}}_{\text{test}}$, the translation quality of θ^{sen} is superior to θ^{doc} and θ^{augdoc} at any chosen value of $\varepsilon > 1$ on both datasets. Without additional training data from WMT22, the private training process of θ^{doc} again becomes unstable for long sequences. This instability results in a divergent loss in the training process and a significant drop in translation quality, even worse with the added noise from DP-SGD. As the base of θ^{augdoc}, $\theta^{\text{augdoc}}_{\text{zero-shot}}$ benefits θ^{augdoc}, which ensures the translation quality to be higher than zero in terms of BLEU even at $\varepsilon = 1$ on both datasets. The document-level training model's results also vary significantly across different values of ε in the private training setups for both datasets.

The privacy/utility trade-off evaluation also shows that the language pair of the dataset has impact on the translation quality of the private training model. To close the gap between the translation quality of θ^{sen} and θ^{augdoc} on $\mathcal{D}^{\text{sen}}_{\text{test}}$, training with more data should be considered.

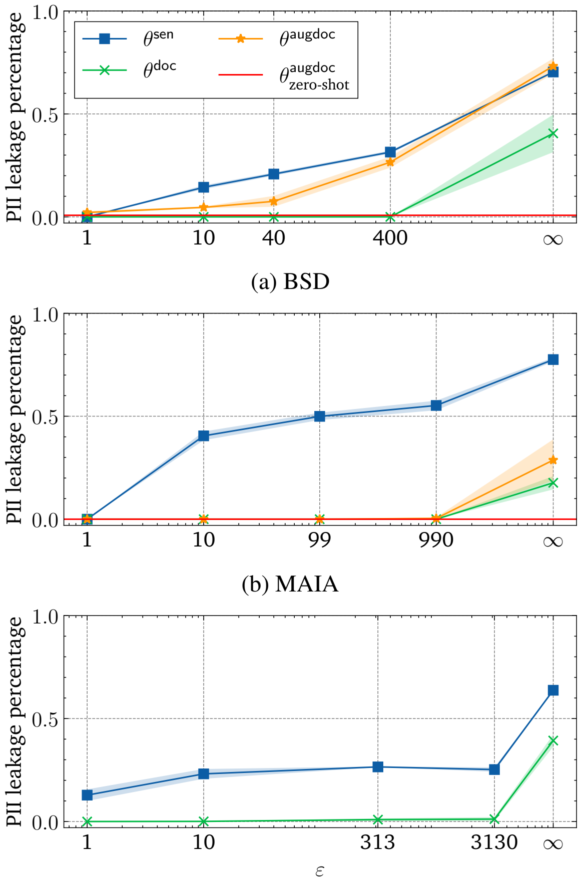
7.2 PII extraction and error analysis

Table 2 shows an example of leakage over three different ε values across three model settings. Overall, without training with DP-SGD, we observe the model to overfit to utterances with PII, even when using θ^{doc} and θ^{augdoc}. By training with DP-SGD, there is no leakage with the document-level models. However, this is not the case for θ^{sen}, even going down to $\varepsilon = 10$, where the customer's name and website URL seem to be memorized. This suggests that θ^{sen} might require a more stringent privacy budget (lower ε) to prevent overfitting to sensitive information compared to document-level models.

8 Conclusion

We have examined the privacy leakage of sentence-level vs. document-level approaches using the loss-based membership inference attack on the mLongT5 model. The results show that a sentence-level model has more risks of privacy leakage than a document-level model. Specifically, a sentence-level model is more likely to overfit compared to a document-level model, which can lead to more confident guessing of sentence-level training instances.

Furthermore, regarding the privacy/utility trade-off in the document-level model, optimizing transformer models for long texts, especially with DP-SGD, is a challenging task that requires more data than our downstream dataset. We demonstrate our solution to this problem with an augmented training technique, using a large public dataset, in our case WMT22, to achieve an acceptable utility, then fine-tune on the downstream dataset to achieve the best privacy/utility trade-off. For future work, we aim to design a better MIA that takes into account the correlation aspect of NLP datasets.



(a) BSD

(b) MAIA

(c) Europarl

Figure 5: PII leakage percentage on $\mathcal{D}^{\text{sen}}_{\text{sampled}}$.

Example	Epsilon	θ^{sen}	θ^{doc}	θ^{augdoc}
	∞	✓	✓	✓
	990	✓	×	×
Agent: Good Morning Dipl.-Ing. Bastian Heuser	10	✓	×	×
	∞	✓	×	×
	990	✓	×	×
Agent: So you would like to cancel the reorder of the lamp?	10	×	×	×
	∞	✓	×	×
	990	✓	×	×
Customer: The standing lamp or the hanging lamp?	10	✓	×	×
	∞	✓	✓	✓
	990	✓	×	×
Agent: - Go to http://www.suessebier.de/	10	✓	×	×
	∞	✓	✓	✓
	990	✓	×	×
Agent: Do you have the order number that starts 160. . . .. ?	10	×	×	×
	∞	✓	✓	✓
	990	✓	×	×
Agent: Thank you - so the Bärer GmbH (140 x 200 cm), Samt in Nachtgrau ?	10	×	×	×

Table 2: Examples of leakage from MAIA $\mathcal{D}^{\text{sen}}_{\text{train}}$ using the MIA and PII evaluation on sentence-level and document-level models. The utterances are collected from within a dialogue. The color depicts the selected PII by `Presidio`. ✓denotes leakage. × denotes no leakage.

Limitations

Although scaling DP to the document level in NLP shows promise, there are several notable pitfalls in our work that need to be addressed in future work:

1. The training data for the document-level scenario is insufficient.
2. Loss-based MIA does not consider the correlation between data and the PII evaluation schema is not perfect.

For point 1, since we are aiming to find a sensitive dataset (e.g. multiple instances of PII), we are limited by the availability of such data for NMT. We are also limited by the size of the document-level dataset when concatenating sentences, which is crucial for training a document-level model. As Zhuocheng et al. (2023b) suggest, we need at least four million training instances to outperform the sentence-level model, while even with the larger Europarl dataset we have around 140k.

Regarding point 2, the loss-based MIA has a significant limitation as it does not take into account the correlation between data. We argue that document-level privacy is stronger than sentence-level privacy when considering this correlation. Therefore, a better MIA method that considers the correlation between data is needed to prove this claim. Future work should investigate the correlation between data and how it affects the privacy guarantee, such as Humphries et al. (2023), but for NLP tasks. In addition, the PII evaluation schema focuses more on risk assessment than strict evaluation. Our PII evaluation relies on the model's confidence in identifying true positive predictions from the MIA and the detection of PII is carried out automatically with `Presidio`, which may result in false positives. Additionally, for the case of NMT, we only consider PII in the target language, assuming the attacker has access to both the source and target instances. This is relevant to the performance of PII detection in source inputs which are not English.

Acknowledgements

This project was supported by the PrivaLingo research grant (Hessisches Ministerium des Innern und für Sport). This work has also been partly supported by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the UA Ruhr (https://uaruhr.de). Thanks to Erion Çano and Lena Held for helpful feedback on a preliminary draft.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. 2022. Large-scale differentially private BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6481–6491, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. 2018. JAX: composable transformations of Python+NumPy programs. `http://github.com/google/jax`.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2280–2292.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Cynthia Dwork and Aaron Roth. 2013. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3-4):211–407.

Ana C Farinha, M. Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José G. C. de Souza, Helena Moniz, and André F. T. Martins. 2022. Findings of the WMT 2022 shared task on chat translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 724–743, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and Increasing Context Usage in Context-Aware Machine Translation. 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 6467–6478, Online. Association for Computational Linguistics.

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.

Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. 2023. Flax: A neural network library and ecosystem for JAX. `http://github.com/google/flax`.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. ArXiv:2302.09210 [cs].

Sorami Hisamoto, Matt Post, and Kevin Duh. 2020. Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system? *Transactions of the Association for Computational Linguistics*, 8:49–63.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2022. Membership Inference Attacks on Machine Learning: A Survey. *ACM Computing Surveys*, 54(11s):1–37.

Lijie Hu, Ivan Habernal, Lei Shen, and Di Wang. 2024. Differentially Private Natural Language Models: Recent Advances and Future Directions. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 478–499, St. Julian's, Malta. Association for Computational Linguistics.

Thomas Humphries, Simon Oya, Lindsey Tulloch, Matthew Rafuse, Ian Goldberg, Urs Hengartner, and Florian Kerschbaum. 2023. Investigating Membership Inference Attacks under Data Dependencies. In *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*, pages 473–488, Los Alamitos, CA, USA. IEEE Computer Society.

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*, pages 13914–13934, Toronto, Canada. Association for Computational Linguistics.

Timour Igamberdiev, Doan Nam Long Vu, Felix Kuennecke, Zhuo Yu, Jannik Holmer, and Ivan Habernal. 2024. DP-NMT: Scalable differentially private machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 94–105, St. Julians, Malta. Association for Computational Linguistics.

Bargav Jayaraman and David Evans. 2019. Evaluating differentially private machine learning in practice. In

28th USENIX Security Symposium (USENIX Security 19), pages 1895–1912, Santa Clara, CA. USENIX Association.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2022. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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*.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. 2022. An empirical analysis of memorization in fine-tuned autoregressive language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1816–1826, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ilya Mironov. 2017. Rényi Differential Privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275, Santa Barbara, CA. IEEE.

Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. Scigen: a dataset for reasoning-aware text generation from scientific tables. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Natalia Ponomareva, Jasmijn Bastings, and Sergei Vassilvitskii. 2022. Training text-to-text transformers with privacy guarantees. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2182–2193, Dublin, Ireland. Association for Computational Linguistics.

Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. 2023. How to dp-fy ml: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation. ArXiv:2304.12959 [cs].

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

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.

Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. 2019. Designing the business conversation corpus. In *Proceedings of the 6th Workshop on Asian Translation*, pages 54–61, Hong Kong, China. Association for Computational Linguistics.

Manuel Senge, Timour Igamberdiev, and Ivan Habernal. 2022. One size does not fit all: Investigating strategies for differentially-private learning across NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7340–7353, Abu Dhabi, UAE.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *2017*

IEEE Symposium on Security and Privacy (SP), pages 3–18. ISSN: 2375-1207.

David Uthus, Santiago Ontanon, Joshua Ainslie, and Mandy Guo. 2023. mLongT5: A multilingual and efficient text-to-text transformer for longer sequences. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9380–9386, Singapore. Association for Computational Linguistics.

Rachel Wicks and Matt Post. 2023. Identifying context-dependent translations for evaluation set production. In *Proceedings of the Eighth Conference on Machine Translation*, pages 452–467, Singapore. Association for Computational Linguistics.

Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. 2023. Document Flattening: Beyond Concatenating Context for Document-Level Neural Machine Translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 448–462, Dubrovnik, Croatia. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282.

Ying Yin and Ivan Habernal. 2022. Privacy-preserving models for legal natural language processing. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 172–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zhang Zhuocheng, Shuhao Gu, Min Zhang, and Yang Feng. 2023a. Addressing the length bias challenge in document-level neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11545–11556, Singapore. Association for Computational Linguistics.

Zhang Zhuocheng, Shuhao Gu, Min Zhang, and Yang Feng. 2023b. Scaling law for document neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8290–8303, Singapore. Association for Computational Linguistics.

A Differential privacy and DP-SGD

Differential privacy (DP) (Dwork and Roth, 2013) is a mathematical framework that ensures that the output of an analysis on a dataset remains unchanged within a specific threshold when any data point is added or removed from the dataset. More formally, for a privacy budget $\varepsilon \geq 0, \delta \in [0, 1]$, a mechanism $\mathcal{M} \colon D^n \to \mathcal{R}^k$ is (ε, δ) differentially private if for all datasets D and D' that differ in at most one instance, and for all $S \subseteq \text{Range}(\mathcal{M})$:

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\varepsilon) \cdot \Pr[\mathcal{M}(D') \in S] + \delta \tag{3}$$

In other words, a mechanism $\mathcal{M}$ is (ε, δ) differentially private if the probability that the mechanism $\mathcal{M}$ returns a response $s \in S$ on dataset D is at most $\exp(\varepsilon)$ times the probability of the mechanism $\mathcal{M}$ returns a response $s \in S$ on dataset D'.

According to the definition, the smaller the privacy budget ε, the greater privacy guarantees the $\mathcal{M}$ mechanism provides, due to its exponential nature, making the differing instance of D and D' indistinguishable. This provides individuals with plausible deniability, as an attacker cannot be certain whether a specific instance belongs to the dataset D or not. However, choosing the appropriate privacy budget ε is crucial to ensure the privacy guarantee of the $\mathcal{M}$ mechanism. If the privacy budget is too large, the $\mathcal{M}$ mechanism will not provide a satisfactory privacy guarantee. On the other hand, if the privacy budget is too low, then the noise added to the query is very high; thus making the mechanism $\mathcal{M}$ impractical. To achieve the desired outcome, a compromise must be made between utility and privacy in the DP application, this is known as the *privacy-utility trade-off* (Dwork and Roth, 2013). Therefore, selecting the appropriate privacy budget for mechanism $\mathcal{M}$ is crucial. In addition, the privacy budget is not fixed and can be adjusted. The privacy budget is adjustable depending on the specific use case, data, and privacy preferences. Typically, to achieve the $\mathcal{M}$ mechanism that satisfies the DP definition, we generally apply noise sampled from the Gaussian distribution to the 'raw' output.

In the case of deep learning, we can apply DP during the training process, in particular prior to the optimization step of a deep neural network, acting as the data analyst, as in the case of DP-SGD (Abadi et al., 2016). This method is presented in Algorithm 1, in which the empirical loss function $L(\theta)$ is minimized with a noisy variant of SGD.

Algorithm 1: Differentially private SGD (Outline) (Abadi et al., 2016)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t \leftarrow \mathbf{g}_t(x_i)/\max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output : θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

During each SGD step t, (Abadi et al., 2016) calculate the gradient $\nabla_\theta \mathcal{L}(\theta, x_i)$ for a random subset of samples L_t via Poisson Sampling. The ℓ_2-norm of each gradient is then clipped, Gaussian noise $\mathcal{N}(0, \sigma^2 C^2 \mathbf{I})$ is added, and the average is taken over all noisy gradients for each element of L_t. A step is then taken in the reverse direction of this noisy gradient to update parameters θ_t. The algorithm aims to prevent over-optimization towards individual data points in the training dataset.

B Data preparation

Figure 6b shows an example of the MAIA dataset with replaced PII data.

MAIA The preprocessing for the MAIA dataset is similar to the BSD dataset. Moreover, since MAIA is a real-world dataset from Unbabel's client, it is anonymized before being released. To make the dataset more realistic, we replace the anonymized PII data with artificial PII data. We use `Faker`[8] to generate fake PII data and replace the pre-anonymized PII data in the MAIA dataset. In each dialogue, we keep the replaced PII data consistent across all utterances (e.g., #NAME# is always replaced by one artificial name within a dialogue). We also use localized fake data for each language (e.g. #PRS_ORG# is replaced by a German company name).

[8] https://faker.readthedocs.io/en/

Notation	Description
$\mathcal{D}^{\textsf{sen}}_{\text{train}}$	Sentence-level training data
$\mathcal{D}^{\textsf{sen}}_{\text{val}}$	Sentence-level validation data
$\mathcal{D}^{\textsf{sen}}_{\text{test}}$	Sentence-level test data
$\mathcal{D}^{\textsf{doc}}_{\text{train}}$	Document-level training data
$\mathcal{D}^{\textsf{doc}}_{\text{val}}$	Document-level validation data
$\mathcal{D}^{\textsf{doc}}_{\text{test}}$	Document-level test data
θ^{sen}	Finetuned model on $\mathcal{D}^{\textsf{sen}}_{\text{train}}$
θ^{doc}	Finetuned model on $\mathcal{D}^{\textsf{doc}}_{\text{train}}$
$\theta^{\text{augdoc}}_{\text{zero-shot}}$	Finetuned model on document-level WMT22
θ^{augdoc}	Finetuned model on $\mathcal{D}^{\textsf{doc}}_{\text{train}}$ with $\theta^{\text{augdoc}}_{\text{zero-shot}}$ checkpoint

Table 3: Notation used in this work.

C Hyperparameters

We first consider the optimal maximum sequence length. For tokenization of the training data, we use the SentencePiece tokenizer (Kudo and Richardson, 2018) that comes with mLongT5. The tokenizer is trained on mC4 (Raffel et al., 2020) with a vocabulary size of 256,384. As shown in Table 4, all datasets have a long tail distribution of token length.

Regarding hyperparameter tuning, we divide this into two cases: (1) Normal training and (2) private training. For both cases, we always set the maximum sequence length to the longest sequence in the training dataset. We conducted an experiment with truncated sequences at 512, 256, and 128 tokens on θ^{doc}. The results of 512 and 256 token sequence lengths are better than setting the model to the longest sequence in the training data, when using a high value of ε. However, the results at small ε values is indifferent to that reported in this work at the longest sequence of the training data.

Table 5 shows the final hyperparameters for each dataset.

C.1 Hyperparameter tuning for normal training

In normal training, we only use one seed for hyperparameter tuning, but three runs for each final selected hyperparameter to get the average performance. We conduct experiments on two H100 GPUs.

		Train			Validation			Test		
		μ	σ^2	max	μ	σ^2	max	μ	σ^2	max
BSD	Japanese	491	165	1007	484	155	843	495	150	1025
	English	499	174	1090	486	163	870	495	158	1060
MAIA	German	589	278	1606	466	440	1101	515	488	1200
	English	555	262	1504	180	174	1034	242	230	1160
Europarl	German	402	397	10899	349	365	7077	351	368	6754
	English	371	368	10046	318	334	6923	320	336	5917

Table 4: Maximum token length, approximate mean and standard variation of each language in the document-level BSD and MAIA datasets. We use the SentencePiece tokenizer to tokenize the data.

Dataset	ϵ	Training Unit	Max. seq. length	lr	Epochs	Total Batch Size
MAIA	∞	Sentence-Level	128	$1e-3$	30	32
	∞	Document-Level	1610	$3e-3$	25	2
	$\{990, 99, 10, 1\}$	Sentence-Level	128	$1e-2$	30	1024
	$\{990, 99, 10, 1\}$	Document-Level	1610	$1e-2$	100	256
BSD	∞	Sentence-Level	64	$1e-3$	15	32
	∞	Document-Level	1100	$3e-3$	30	2
	$\{400, 40, 10, 1\}$	Sentence-Level	64	$1e-2$	15	1024
	$\{400, 40, 10, 1\}$	Document-Level	1100	$1e-2$	100	512
Europarl	∞	Sentence-Level	128	$1e-4$	16	128
	∞	Document-Level	1500	$1e-4$	30	128
	$\{3130, 313, 10, 1\}$	Sentence-Level	128	$1e-3$	16	1,048,576
	$\{3130, 313, 10, 1\}$	Document-Level	1500	$1e-3$	100	131,072
WMT22-JA-EN	∞	Document-Level	1200	$1e-2$	2	4
WMT22-DE-EN	∞	Document-Level	1500	$1e-2$	2	4

Table 5: Final results for hyperparameter search.

Dataset	# Member	# Non-member
BSD	4147	4147
MAIA	4597	4597
Europarl	363,538	363,538

(a) Number of examples used for sentence-level loss-based MIA.

Dataset	# PII
BSD	1286
MAIA	1156
Europarl	27,527

(b) Number of PII in $\mathcal{D}^{\mathsf{sen}}_{\mathtt{sampled}}$ of each dataset

Table 6: Statistics of datasets used for MIA evaluation.

Lang. Pair	Sen.-level	Doc.-level
JA-EN	33,875,119	851,525
DE-EN	295,805,439	4,779,636

Table 7: Number of training examples in WMT22 dataset.

C.2 Hyperparameter tuning for private training

For the final results with DP-SGD, we run two trials for each hyperparameter configuration, with five different seeds for each final selected hyperparameter, for each ε value. We use the same notation for *epochs* when running DP-SGD training with Poisson sampling as Abadi et al. (2016), being $\frac{N}{L}$. Next, we utilize very large batch sizes for both of these methods, setting L to a large value and building up the resulting drawn batches with gradient accumulation. All private training experiments are conducted using one H100 GPU, due to the limitation of duplicating examples in lots when sampling to multiple GPUs. Similar to normal training, we keep the same maximum sequence length for both datasets in private training.

```
{
...
"de": Hallo, können sie mir sagen wann das bestellte Bett ca Versand wird?
"en": Hello, Can you tell me when the ordered bed will be approximately shipped?
"de": Guten Morgen #NAME#,
"en": Good Morning #NAME#.
"de": vielen Dank, dass Sie #PRS_ORG# kontaktiert haben. Ich hoffe, dass es Ihnen gut geht.
"en": Thank you for contacting #PRS_ORG# I hope you are well.
...
}
```

(a) Sentence-level original training pairs

```
{
...
"de": Kunde: Hallo, können sie mir sagen wann das bestellte Bett ca Versand wird?
"en": Customer: Hello, Can you tell me when the ordered bed will be approximately shipped?
"de": Agent: Guten Morgen Olav Kusch,
"en": Agent: Good Morning Olav Kusch.
"de": Agent: vielen Dank, dass Sie Hethur Ullmann GmbH & Co. KG kontaktiert haben. Ich hoffe, dass es Ihnen gut geht.
"en": Agent: Thank you for contacting Hethur Ullmann GmbH & Co. KG I hope you are well.
...
}
```

(b) Sentence-level with artificial replaced PII training pairs

```
{
"de": ... Kunde: Hallo, können sie mir sagen wann das bestellte Bett ca Versand wird? Agent: Guten Morgen Olav
Kusch Agent: vielen Dank, dass Sie Hethur Ullmann GmbH & Co. KG kontaktiert haben. Ich hoffe, dass es Ihnen gut geht. ...
"en": ... Customer: Hello, Can you tell me when the ordered bed will be approximately shipped? Agent: Good Morning Olav
Kusch. Agent: Thank you for contacting Hethur Ullmann GmbH & Co. KG I hope you are well. ...
}
```

(c) Document-level with artificial replaced PII training pair (utterances within a dialogue)

Figure 6: Difference between training examples for the document-level vs sentence-level MAIA dataset.

C.3 Augmented training for document-level models

We then fine-tune the mLongT5 checkpoint on the concatenated documents for 2 epochs with a batch size of 16 on two H100 GPUs. The hyperparameter search space is the same as for normal training, as described above. For Japanese to English translation, we use the entire WMT dataset. However, for German to English translation, we only use the first part of the dataset. This is done to ensure that the dataset size is equal to that of the Japanese to English dataset (851,525), due to time constraints.[9]

D BERTScore Modification

We also use BERTScore (Zhang et al., 2020) for semantic similarity evaluation. BERTScore uses RoBERTa embeddings (Liu et al., 2019) to compute the similarity between the candidate translation and the reference translation. However, its embeddings are limited to 512 tokens, which is not enough for our task. Therefore, we modify BERTScore to use Longformer[10] embeddings (Beltagy et al., 2020) instead of RoBERTa embeddings. Longformer's embeddings are able to encode long sentences up to 4,096 tokens. Another modification that we make to BERTScore in this work is rescaling the score baseline to make it more readable.

The score is computed from the seventh layer output of Longformer's embeddings, since it has the best correlation with human judgment on the WMT16 Metrics Shared Task (Zhang et al., 2020). BERTScore uses pre-normalized vectors for cosine similarity, resulting in computed scores the range $[-1, 1]$. However, in practice, the observed BERTScore values are often limited to a narrow range (Moosavi et al., 2021; Zhang et al., 2020). For instance, when using the default large RoBERTa[11] model, BERTScore typically falls between 0.85 and 0.95. This is due to the learned geometry of contextual embeddings which results in different scores from different embeddings. Although this characteristic does not affect BERTScore's ability to rank text generation systems, it does make the resulting score less comprehensible to humans. To address this issue, Zhang et al. (2020) rescale BERTScore using its empirical lower bound b as a baseline. The computation of b is carried out using Common Crawl[12] monolingual datasets. For each language and contextual embed-

[9]It takes 98 hours to finish one epoch on the entire document-level WMT22. We also split the dataset by half and trained the model on each part. The results are poor compared to using only a small part of the data, despite training of each part to simulate training the entire dataset in one epoch.

[10]https://huggingface.co/allenai/longformer-base-4096

[11]https://huggingface.co/roberta-large

[12]https://commoncrawl.org/

```
...
{
”ja_speaker”: 土井さん
”ja_sentence”: 稲田さん、H社の高市様からお電話です。
”en_speaker”: Doi−san
”en_sentence”: Inada−san, you have a call from Mr. Takaichi of Company H.
}
{
”ja_speaker”: 稲田さん
”ja_sentence”: もしもし、稲田です。
”en_speaker”: Inada−san
”en_sentence”: Hello, this is Inada.
}
...
```

(a) Sentence-level original training pairs

```
{
...
”ja”: 土井さん: 稲田さん、H社の高市様からお電話です。
”en”: Doi−san: Inada−san, you have a call from Mr. Takaichi of Company H.
”ja”: 稲田さん: もしもし、稲田です。
”en”: Inada−san: Hello, this is Inada.
...
}
```

(b) Sentence-level modified training pairs

```
{
”ja”: ... 土井さん: 稲田さん、H社の高市様からお電話です。 稲田さん: もしもし、稲田です。...
”en”: ... Doi−san: Inada−san, you have a call from Mr. Takaichi of Company H. Inada−san: Hello, this is Inada. ...
}
```

(c) Document-level training pair (utterances within a dialogue)

Figure 7: Difference between training examples for the document-level vs sentence-level BSD dataset.

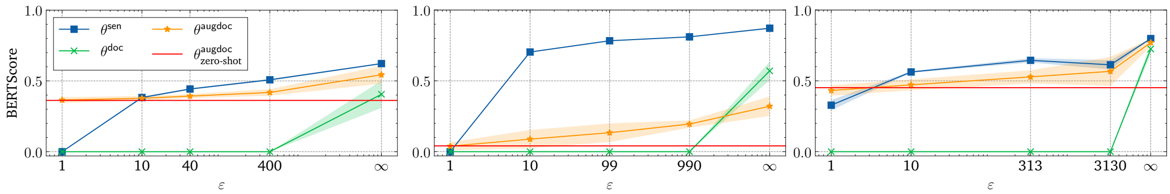
```
...
{
”de”: Günter Gloser: Die Fähigkeit des Menschenrechtsrats, sein Mandat zu erfüllen,
ist untrennbar mit seiner Zusammensetzung verbunden.
”en”: Günter Gloser: The capacity of the Human Rights Council to
fulfil its mandate is inextricably linked to its composition.
}
{
”de”: Günter Gloser: Lassen Sie mich auch hier daran erinnern, dass die Generalversammlung
der Vereinten Nationen von den Mitgliedstaaten des Rats die Einhaltung höchster Menschenrechtsstandards erwartet.
”en”: Günter Gloser: Allow me to recall once more that the General Assembly of the United Nations
expects the member nations of the Council to observe the highest standards of human rights.
}
...
```

(a) Sentence-level original training pairs

```
{
”de”: Günter Gloser: ...
Die Fähigkeit des Menschenrechtsrats, sein Mandat zu erfüllen,
ist untrennbar mit seiner Zusammensetzung verbunden.
Lassen Sie mich auch hier daran erinnern, dass die Generalversammlung der Vereinten Nationen
von den Mitgliedstaaten des Rats die Einhaltung höchster Menschenrechtsstandards erwartet.
...
”en”: Günter Gloser: ...
The capacity of the Human Rights Council to
fulfil its mandate is inextricably linked to its composition.
Allow me to recall once more that the General Assembly of the United Nations
expects the member nations of the Council to observe the highest standards of human rights.
...
}
```

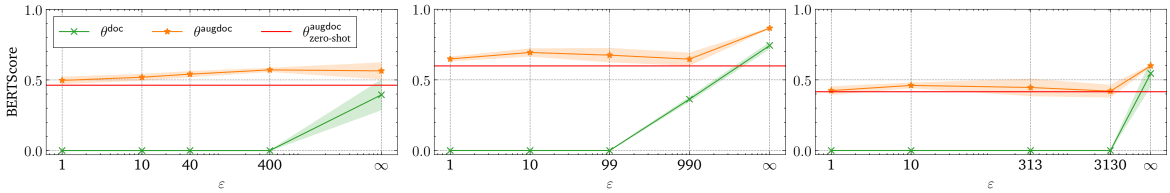
(b) Document-level training pair (utterances within a speech)

Figure 8: Difference between training examples for the document-level vs sentence-level Europarl dataset.



(a) BSD (b) MAIA (c) Europarl

Figure 9: BERTScore on $\mathcal{D}^{\mathsf{sen}}_{\mathtt{test}}$ for all four model fine-tuning configurations. Lower ε corresponds to better privacy.



(a) BSD (b) MAIA (c) Europarl

Figure 10: BERTScore on $\mathcal{D}^{\text{doc}}_{\text{test}}$ for the three document-level model fine-tuning configurations. Lower ε corresponds to better privacy.

ding model, one million candidate-reference pairs are created by grouping two random sentences. Due to the random pairing and corpus diversity, each pair has very low lexical and semantic overlap (BLEU computed on these pairs is around zero). To compute the value of b, Zhang et al. (2020) take the average of BERTScore computed on the sentence. After that, using baseline b, we get a linearly rescaled BERTScore.

$$\hat{F}_{\text{BERT}} = \frac{F_{\text{BERT}} - b}{1 - b}$$

The result $\hat{F}_{\text{BERT}}$ typically ranges between 0 and 1, with anything below this range clipped to 0. As Zhang et al. (2020) note, this method does not affect the ranking ability or human correlation of BERTScore, as measured by Pearson's and Kendall's coefficients, but to enhance the readability of the score. Since the Longformer rescaled BERTScore is not available, we compute it ourselves.

E MIA experiment details

Table 6a shows the number of members and non-members for MIA evaluation.

F Privacy/utility trade-off evaluation with BERTScore

Apart from using BLEU for evaluating the privacy/utility trade-off (Section 5.5), we also use BERTScore for the evaluation of translation quality at the document level, with respect to semantic similarity.

F.1 Privacy/utility trade-off on $\mathcal{D}^{\text{doc}}_{\text{test}}$

Figure 10 shows the BERTScore of θ^{doc} and θ^{augdoc} on $\mathcal{D}^{\text{doc}}_{\text{test}}$. It is evident that the BERTScore decreases as the value of ε is decreased for both datasets, similar to the BLEU score results. Compared to BLEU scores, the main difference is that θ^{augdoc} is more stable across different values of ε, in particular for the BSD dataset. It is also worth noting that at $\varepsilon = 400$ on BSD, the BERTScore of θ^{doc} is 0, while the BLEU score is still around 2. On MAIA, the BERTScore of θ^{doc} and θ^{augdoc} shows less variation at $\varepsilon = \infty$ compared to the BLEU score. Regarding Europarl, the BERTScore of θ^{doc} is 0 when training with DP-SGD, while θ^{augdoc} performance trend is similar to BSD and MAIA results.

F.2 Privacy/utility trade-off on $\mathcal{D}^{\text{sen}}_{\text{test}}$

Figure 9 shows the BERTScore of θ^{sen}, θ^{doc} and θ^{augdoc} on $\mathcal{D}^{\text{sen}}_{\text{test}}$.

MAIA On MAIA, the BERTScore of θ^{sen} slowly decreases as the value of ε is decreased, for about 0.1 points per decrease in ε from ∞ to 10. It is interesting that the BERTScore of θ^{augdoc} varies a lot across different values of ε, e.g., 0.088 ± 0.079 at $\varepsilon = 10$ and 0.134 ± 0.081 at $\varepsilon = 99$.

BSD On BSD, the BERTScores of θ^{sen}, θ^{doc} and θ^{augdoc} are very close to one other, especially at $\varepsilon = \infty$. From $\varepsilon = 400$, we no longer observe any meaningful translation quality from θ^{doc}. The BERTScores of θ^{sen} and θ^{augdoc} start to converge and become equal at $\varepsilon = 10$, which demonstrates that the semantic content of θ^{sen} and θ^{augdoc} equals at the sentence level on the BSD dataset. As expected, the BERTScore of θ^{sen} at $\varepsilon = 1$ is equal to zero due to the high amount of added noise during the training process with DP-SGD, while the BERTScore of θ^{augdoc} is equal to that of $\theta^{\text{augdoc}}_{\text{zero-shot}}$ at approximately 0.38.

Europarl On Europarl, the BERTScores of θ^{sen}, θ^{doc} and θ^{augdoc} are even closer at $\varepsilon = \infty$ than on BSD. While at $\varepsilon = 3130$ θ^{doc} fails to generate a sentence, θ^{sen} and θ^{augdoc} have similar drops in their performance. Thanks to the large amount of training data, θ^{sen} keeps the results at 0.4 BERTScore with $\varepsilon = 1$. Unlike the BLEU score results, θ^{sen} impressively maintains the semantic similarity of the generated sentence with 30% reduction.

G Discussion on translation quality

Finally, Table 8 and Table 9 show that the document-level training model mainly duplicates the sentence until it reaches the maximum sequence length. Overall, θ^{augdoc} shows better performance for private training than θ^{doc} and is close to θ^{sen} performance with post-processing.

H DP guarantees in our experiments

To provide all the information needed to understand our privacy guarantees, we follow the guidelines outlined in Ponomareva et al. (2023).

1. **DP setting.** We provide a central DP guarantee where the service provider is trusted to correctly implement the mechanism.

2. **Instantiating the DP Definition**

    (a) *Data accesses covered*: Our DP guarantees apply only to a single training run. We don't account for hyperparameter tuning in our guarantees. Public multilingual C4 data (Raffel et al., 2020; Xue et al., 2021) is used for pre-training mLongT5.

    (b) *Final mechanism output*: Only the model predictions, such as the translated sentences generated by the models trained with DP, are released. The mechanism's output is technically the full sequence of privatized gradients, and the guarantee also applies at this level. Hence, all checkpoints are protected and can be released publicly.

    (c) *Unit of privacy.* Since we are working in the NLP context, we consider sentences and documents as the unit of privacy. The sentence-level unit is an utterance in a conversation, typically a single sentence with a maximum length of 64 to 128 tokens, depending on the dataset. The document-level unit is the whole conversation dialogue, which can be composed of multiple sentences. Thus, the maximum length of the document is not limited, and in our experiments, the maximum length of the document is up to 1,700 tokens. Token counting is done after tokenization using SentencePiece (Kudo and Richardson, 2018). We demonstrate in our experiments that sentence-level privacy is weaker than document-level privacy. However, group privacy can be used to achieve document-level privacy from sentence-level privacy.

    (d) *Adjacency definition for "neighboring" datasets*: We use the add-or-remove adjacency definition.

3. (a) *Type of accounting used*: RDP-based accounting.

    (b) *Accounting assumptions*: We correctly use Poisson sampling.

    (c) *The formal DP statement*: We use various levels of ε values: $1, 10, 40, 99, 400, 990$. Our δ is set to 10^{-8}.

    (d) *Transparency and verifiability*: We are going to open source our code based on the open-source DP-NMT framework (Igamberdiev et al., 2024).

Model	ϵ	System Output
θ^{sen}	∞	Customer: I just bought a book with my Geisler Conradi GmbH, it seems to be on it.
	990	Customer: I bought a book directly with my Geisler Conradi GmbH, it seems to be on it.
	99	Customer: I bought a book directly with my Geisler Conradi GmbH, it seems to be on it.
	10	Customer: I bought the Geisler Conradi GmbH, a book it seems to be on.
θ^{doc}	∞	Customer: I bought a book directly with my Geisler Conradi GmbH, it seems to be on it.
	990	Customer: I ordered directly with Geisler Conradi GmbH, a book. It seems to be on.. \n Customer: I ordered directly with my Geisler Conradi GmbH, a book. It seems to be on.. \n Customer: I ordered directly with my Geisler Conradi GmbH, a book. It seems to be on.. \n Customer: I ordered directly with my Geisler Conradi GmbH, a book. It seems to be on.. \n Customer: I ordered directly with my Geisler Conradi GmbH, a book. It seems to be on..
θ^{augdoc}	∞	Customer: I bought a book directly with my Geisler Conradi GmbH and it seems to be on it.
	990	Customer: I bought directly with my Geisler Conradi GmbH, a book it seems to be on it.␣ Customer: I bought directly with my Geisler Conradi GmbH, a book it seems to be on it.␣ Customer: I bought directly with my Geisler Conradi GmbH, a book it seems to be on it.␣ Customer: I bought directly with my Geisler Conradi GmbH, a book it seems to be on it.␣ Customer: I bought directly with my Geisler Conradi GmbH, a book it seems to be on it.
	99	Customer: Have directly with my Geisler Conradi GmbH, bought a book it seems to be on it too.␣ Customer: Have directly with my Geisler Conradi GmbH, bought a book it seems to be on it too.␣ Customer: Have directly with my Geisler Conradi GmbH, bought a book it seems to be on it too.␣ Customer: Have directly with my Geisler Conradi GmbH, bought a book it seems to be on it too.␣ Customer: Have directly with my Geisler Conradi GmbH, bought a book it seems to be on it too.
	10	Customer: Have directly with my Geisler Conradi GmbH, a book bought it seems to be on it too. Customer: Have directly with my Geisler Conradi GmbH, a book bought it seems to be on it too.␣ Customer: Have directly with my Geisler Conradi GmbH, a book bought it seems to be on it too.␣ Customer: Have directly with my Geisler Conradi GmbH, a book bought it seems to be on it too. Customer: Have directly with my Geisler Conradi GmbH, a book bought it seems to be on it too.
	1	Customer: Have directly with my Geisler Conradi GmbH, a book bought it seems to be on it.␣ Customer: Have directly with my Geisler Conradi GmbH, a book bought it seems to be on it.␣ Customer: Have directly with my Geisler Conradi GmbH, a book bought it seems to be on it.␣ Customer: Have directly with my Geisler Conradi GmbH, a book bought it seems to be on it.␣ Customer: Have directly with my Geisler Conradi GmbH, a book bought it seems to be on it.
$\theta^{\text{augdoc}}_{\text{zero-shot}}$	-	Customer: Have directly with my Geisler Conradi GmbH, a book bought it seems to be on it too.␣ Customer: Have directly with my Geisler Conradi GmbH, a book bought it seems to be on it too.␣ Customer: Have directly with my Geisler Conradi GmbH, a book bought it seems to be on it too.␣ Customer: Have directly with my Geisler Conradi GmbH, a book bought it seems to be on it too.␣ Customer: Have directly with my Geisler Conradi GmbH, a book bought it seems to be on it too.
Original	Reference	Customer: I bought a book directly with my Geisler Conradi GmbH, it seems to be on it too.
	Source	Kunde: Habe direkt mit meinem Geisler Conradi GmbH, ein buch gekauft es scheint auch drauf zu sein.

Table 8: Translation sample from $\mathcal{D}^{\mathsf{sen}}_{\text{test}}$ of MAIA dataset

Model	ϵ	System Output
θ^{sen}	∞	Akiyama-san: Oki-san, can you prepare some designs for cups and cards by next week's meeting?
	400	Ms. Murayama: How can you prepare the design of your cup and card for the next meeting?
	40	Mr. Yamamoto-san: How can you prepare your team for the upcoming meeting? \n Michi
	10	Mr. Miyamoto-san: I'm going to make a card for the meeting tomorrow.
θ^{doc}	∞	Mr. Maeda: How many cups and card designs do you need to prepare for next week's meeting? Mr. Maeda: How many cards do you need to design for your team?
	400	秋山さん, 2020-08-20 2020-08-20 2020-08-20 2020-08-20 2020-08-20
θ^{augdoc}	∞	Akiya-san: Bessho-san, can you team up with your team to draft some cup and card designs by next week's meeting?
	400	Mr. Akiya: Mr. Bessho, can you prepare some cup and card design ideas for your team by the meeting next week? Mr. Akiya: Mr. Bessho, can you prepare some cup and card design ideas for your team by the meeting next week? Mr. Akiya:
	40	Mr. Akiya: Mr. Bessho, can you prepare some cup and card design ideas for your team by the meeting next week? Mr. Akiya: Mr. Bessho, can you prepare some cup and card design ideas for your team by the meeting next week? Mr. Akiya:
	10	Mr. Akiyama: Mr. Bessho, can your team prepare some cup and card design ideas by next week's meeting? Mr. Akiyama: Mr. Bessho, can your team prepare some cup and card design ideas by next week's meeting? Mr. Akiyama
	1	Akiyama: Mr. Bessho, can your team prepare some cup and card design ideas by next week's meeting? Mr. Akiyama: Mr. Bessho, can your team prepare some cup and card design ideas by next week's meeting? Mr. Akiyama: Mr. Bess
$\theta^{\text{augdoc}}_{\text{zero-shot}}$	-	Mr. Akiyama: Mr. Bessho, can you prepare some cup and card design ideas for your team by next week's meeting? Mr. Akiyama: Mr. Bessho, can you prepare some cup and card design ideas for your team by next week's meeting? Mr
	Reference	Mr. Akiyama: Ms. Bessho, can your team prepare a few design ideas for the cup and card by next week's meeting?
	Source	秋山さん: 別所さん、来週のミーティングまでにあなたのチームでカップとカードのデザイン案を幾つか準備してもらえますか？

Table 9: Translation sample from $\mathcal{D}^{\mathsf{sen}}_{\mathtt{test}}$ of BSD dataset

Model	ϵ	System Output
θ^{sen}	∞	Hager: In principle, the Freedom Party welcomes all the measures and steps to be taken, both in the EU and in the Member States, to combat money laundering.
	3130	Hager: The freedoms welcome all measures and steps that they are taking in the EU as well as in Member States to combat the money washing. The freedoms welcome all measures and steps that they are taking in the EU as well as in Member States to combat the money washing. The freedoms welcome all measures and steps that they are taking in the EU as well as in Member States to combat the money washing. The freedoms welcome all measures and steps that they are taking in the EU as well as in Member States to combat the money washing
	313	Hager: The freedoms welcome all measures and steps that are taken in the EU as well as in Member States to combat the fraud.
	10	Hager: The liberty welcomes all measures and steps, the EU and the Member States are implementing.
	1	Hager: The Freedom of welcomes all measures and steps, which we have in the EU as well as in the Member States of the Member States to prevent the money-washing in the Member States and the Member States of the EU, and the Member States of the Member States of the EU to prevent the money-washing in the Member States. Hager: The Freedom of welcomes all measures and steps, which we have in the EU as well as in the Member States of the Member States of the EU to prevent the money-washing in the Member States.
θ^{doc}	∞	Hager: The freedoms generally welcome all the measures and steps that are being taken in both the EU and the Member States to combat money laundering.
	3130	\u0110\u1ea3ng: Hager:
	313	Kanker: I would like to thank you for your support. Kanker: I would like to thank you for your support. Kanker: I would like to thank you for your support. Kanker: I would like to thank you for your support. Kanker: I would like to thank you for your support. Kanker: I would like to thank you for your support. Kanker: I would like to thank you for your support. Kanker: I would like to thank you for your support. Kanker: I would like to thank you for your support. Kanker: I would like to thank you for
θ^{augdoc}	∞	Hager: The freedom fighters welcome in principle all the measures and steps taken to combat money laundering, both in the EU and in the Member States.
	3130	Hager: In principle, the freedom-seekers welcome all the measures and steps that are being taken both in the EU and in the Member States to combat money laundering.
	313	Hager: The freedom-seekers welcome in principle all the measures and steps taken both in the EU and in the Member States to combat money laundering. Hager: The freedom-seekers welcome in principle all the measures and steps taken both in the EU and in the Member States to combat money laundering. Hager: The freedom-seekers welcome in principle all the measures and steps taken both in the EU and in the Member States to combat money laundering.
	10	Hager: The freedom-seekers generally welcome all the measures and steps taken both in the EU and in the Member States to combat money laundering. Hager: The freedom-seekers generally welcome all the measures and steps taken both in the EU and in the Member States to combat money laundering. Hager: The freedom-seekers generally welcome all the measures and steps taken both in the EU and in the Member States to combat money laundering. Hager: The freedom-seekers generally welcome all the measures and steps taken both in the EU and in the Member States to combat money laundering
	1	Hager: The freedom-seekers generally welcome all measures and steps taken both in the EU and in the Member States to combat money laundering. Hager: The freedom-seekers generally welcome all measures and steps taken both in the EU and in the Member States to combat money laundering. Hager: The freedom-seekers generally welcome all measures and steps taken both in the EU and in the Member States to combat money laundering. Hager: The freedom-seekers generally welcome all measures and steps taken both in the EU and in the Member States to combat money laundering. Hager:
$\theta^{\text{augdoc}}_{\text{zero-shot}}$	-	Hager: The freedom-seekers generally welcome all measures and steps taken both in the EU and in the Member States to combat money laundering. Hager: The freedom-seekers generally welcome all measures and steps taken both in the EU and in the Member States to combat money laundering. Hager: The freedom-seekers generally welcome all measures and steps taken both in the EU and in the Member States to combat money laundering. Hager: The freedom-seekers generally welcome all measures and steps taken both in the EU and in the Member States to combat money laundering. Hager:
Original	Reference	Hager: The non-attached Members welcome on principle all measures and steps taken both in the EU and in the Member States to combat money laundering.
	Source	Hager: Die Freiheitlichen begrüßen grundsätzlich alle Maßnahmen und Schritte, die man sowohl in der EU als auch in den Mitgliedstaaten zur Bekämpfung der Geldwäsche setzt.

Table 10: Translation sample from $\mathcal{D}^{\text{sen}}_{\text{test}}$ of Europarl dataset