Membership Inference Attacks against Language Models via Neighbourhood Comparison

Justus Mattern[1], **Fatemehsadat Mireshghallah**[2], **Zhijing Jin**[3,4],
Bernhard Schölkopf[3], **Mrinmaya Sachan**[4], **Taylor Berg-Kirkpatrick**[2]
RWTH Aachen[1], UC San Diego[2], MPI for Intelligent Systems[3], ETH Zürich[4]
Correspondence: justusmattern27@gmail.com

Abstract

Membership Inference attacks (MIAs) aim to predict whether a data sample was present in the training data of a machine learning model or not, and are widely used for assessing the privacy risks of language models. Most existing attacks rely on the observation that models tend to assign higher probabilities to their training samples than non-training points. However, simple thresholding of the model score in isolation tends to lead to high false-positive rates as it does not account for the intrinsic complexity of a sample. Recent work has demonstrated that reference-based attacks which compare model scores to those obtained from a reference model trained on similar data can substantially improve the performance of MIAs. However, in order to train reference models, attacks of this kind make the strong and arguably unrealistic assumption that an adversary has access to samples closely resembling the original training data. Therefore, we investigate their performance in more realistic scenarios and find that they are highly fragile in relation to the data distribution used to train reference models. To investigate whether this fragility provides a layer of safety, we propose and evaluate neighbourhood attacks, which compare model scores for a given sample to scores of synthetically generated neighbour texts and therefore eliminate the need for access to the training data distribution. We show that, in addition to being competitive with reference-based attacks that have perfect knowledge about the training data distribution, our attack clearly outperforms existing reference-free attacks as well as reference-based attacks with imperfect knowledge, which demonstrates the need for a reevaluation of the threat model of adversarial attacks.

1 Introduction

The public release and deployment of machine learning models trained on potentially sensitive user data introduces a variety of privacy risks: While embedding models have been shown to leak personal attributes of their data (Song and Raghunathan, 2020), generative language models are capable of generating verbatim repetitions of their training data and therefore exposing sensitive strings such as names, phone numbers or email-addresses (Carlini et al., 2021b). Another source of risk arises from membership inference attacks (MIAs) (Shokri et al., 2016), which enable adversaries to classify whether a given data sample was present in a target model's training data or not. Due to their simplicity and the fact that MIAs are an important component of more sophisticated attacks such as extraction attacks (Carlini et al., 2021b), they have become one of the most widely used tools to evaluate data leakage and empirically study the privacy of machine learning models (Murakonda and Shokri, 2020; Song and Marn, 2020).

Typically, membership inference attacks exploit models' tendency to overfit their training data and therefore exhibit lower loss values for training members (Yeom et al., 2018; Sablayrolles et al., 2019). A highly simple and commonly used baseline attack is therefore the LOSS attack (Yeom et al., 2018), which classifies samples as training members if their loss values are below a certain threshold. While attacks of this kind do generally reap high accuracies, Carlini et al. (2021a) point out a significant flaw: *Good accuracies for attacks of this kind are primarily a result of their ability to identify non-members rather than training data members*, which does arguably not pose important privacy risks. This shortcoming can be attributed to the fact that certain samples such as repetitive or very simple short sentences are naturally assigned higher probabilities than others (Fan et al., 2018; Holtzman et al., 2020), and the influence of this aspect on the obtained model score largely outweighs the influence of a model's tendency to overfit its training samples (Carlini et al., 2021a). To account for this, previous work has introduced the idea of *difficulty calibration mechanisms* (Long
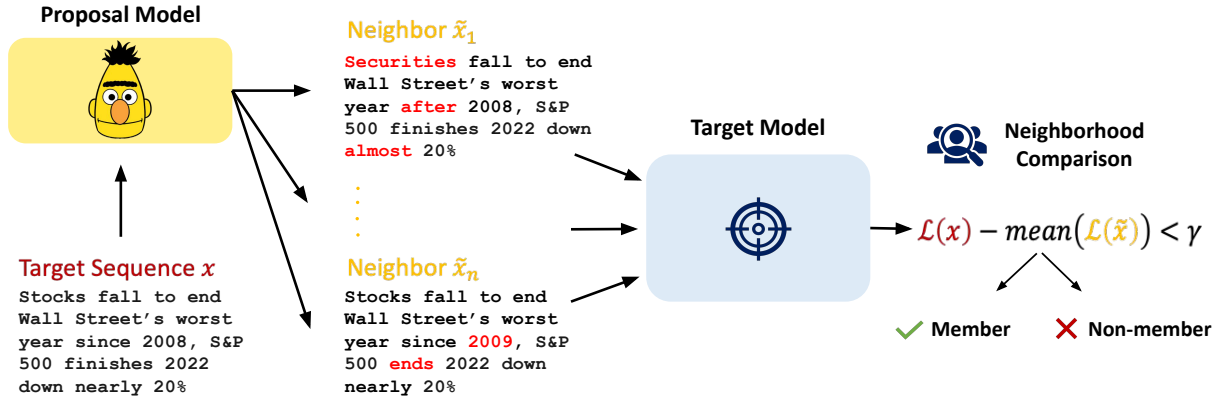
11330

Proposal Model

Target Sequence x
```
Stocks fall to end
Wall Street's worst
year since 2008, S&P
500 finishes 2022
down nearly 20%
```

Neighbor $\tilde{x}_1$
```
Securities fall to end
Wall Street's worst
year after 2008, S&P
500 finishes 2022 down
almost 20%
```

Neighbor $\tilde{x}_n$
```
Stocks fall to end
Wall Street's worst
year since 2009, S&P
500 ends 2022 down
nearly 20%
```

Target Model

Neighborhood Comparison

$$\mathcal{L}(x) - mean\big(\mathcal{L}(\tilde{x})\big) < \gamma$$

✓ **Member**   ✗ **Non-member**

Figure 1: Overview of our attack: Given a target sample x, we use a pretrained masked language model to generate highly similar neighbour sentences through word replacements. Consequently, we compare our neighbours' losses and those of the original sample under the target model by computing their difference. As our neighbours are highly similar to the target sequence, we expect their losses to be approximately equal to the target model and only to be lower if the target sequence was a sample of the model's training data. In this case, the difference should be below our threshold value γ.

et al., 2018; Watson et al., 2022), which aim to quantify the intrinsic complexity of a data sample (i.e., how much of an outlier the given sample is under the probability distribution of the target model) and subsequently use this value to regularize model scores before comparing them to a threshold value.

In practice, difficulty calibration is mostly realized through *Likelihood Ratio Attacks (LiRA)*, which measure the difficulty of a target point by feeding it to *reference models* that help provide a perspective into how likely that target point is in the given domain (Ye et al., 2022; Carlini et al., 2021a; Watson et al., 2022; Mireshghallah et al., 2022a,b). In order to train such reference models, LiRAs assume that an adversary has knowledge about the distribution of the target model's training data and access to a sufficient amount of samples from it. We argue that this is a highly optimistic and in many cases unrealistic assumption: as also pointed out by Tramèr et al. (2022), in applications in which we care about privacy and protecting our models from leaking data (e.g. in the medical domain), high-quality, public in-domain data may not be available, which renders reference-based attacks ineffective. Therefore, we aim to design an attack which does not require any additional data: For the design of our proposed *neighborhood attack*, we build on the intuition of using references to help us infer membership, but instead of using reference models, we use *neighboring samples*, which are textual samples crafted through data augmentations such as word replacements to be non-training members that are as similar as possible to the target point and therefore practically interchangeable with it in almost any context. With the intuition that neighbors should be assigned equal probabilities as the original sample under any plausible textual probability distribution, we then compare the model scores of all these neighboring points to that of the target point and classify its membership based on their difference. Similar to LiRAs, we hypothesize that if the model score of the target data is similar to the crafted neighbors, then they are all plausible points from the distribution and the target point is not a member of the training set. However, if a sample is much more likely under the target model's distribution than its neighbors, we infer that this could only be a result of overfitting and therefore the sample must be a part of the model's training data.

We conduct extensive experiments measuring the performance of our proposed neighborhood attack, and particularly compare it to reference-based attacks with various different assumptions about knowledge of the target distribution and access to additional data. Concretely, amongst other experiments, we simulate real-world reference-based attacks by training reference models on external datasets from the same domain as the target model's training data. We find that neighbourhood attacks outperform LiRAs with more realistic assumptions about the quality of accessible data by up to 100%, and even show competitive performance when we assume that an attacker has perfect knowledge about the target distribution and access to a large amount of high-quality samples from it.

2 Membership Inference Attacks via Neighbourhood Comparison

In this section, we provide a detailed description of our attack, starting with the general idea of comparing neighbouring samples and following with a technical description of how to generate such neighbors.

2.1 General Idea

We follow the commonly used setup of membership inference attacks in which the adversary has grey-box access to a machine learning model f_θ trained on an unknown dataset $\mathcal{D}_{\text{train}}$, meaning that they can obtain confidence scores and therefore loss values from f_θ, but no additional information such as model weights or gradients. The adversary's goal is to learn an attack function $A_{f_\theta} : \mathcal{X} \rightarrow \{0, 1\}$, which determines for each x from the universe of textual samples $\mathcal{X}$ whether $x \in D_{\text{train}}$ or $x \notin D_{\text{train}}$. As mentioned in the previous section, the LOSS attack (Yeom et al., 2018), one of the most simple forms of membership inference attacks, classifies samples by thresholding their loss scores, so that the membership decision rule is:

$$A_{f_\theta}(x) = \mathbb{1}[\mathcal{L}(f_\theta, x) < \gamma]. \qquad (1)$$

More recent attacks follow a similar setup, but perform difficulty calibration to additionally account for the intrinsic complexity of the sample x under the target distribution and adjust its loss value accordingly. Concretely, given a function $d : \mathcal{X} \rightarrow \mathbb{R}$ assigning difficulty scores to data samples, we can extend the the decision rule to

$$A_{f_\theta}(x) = \mathbb{1}[\mathcal{L}(f_\theta, x) - d(x) < \gamma]. \qquad (2)$$

Likelihood Ratio Attacks (LiRAs) (Ye et al., 2022), the currently most widely used form of membership inference attacks, use a sample's loss score obtained from some reference model f_ϕ as a difficulty score, so that $d(x) = \mathcal{L}(f_\phi, x)$. However, this makes the suitability of the difficulty score function dependent on the quality of reference models and therefore the access to data from the training distribution. We circumvent this by designing a different difficulty calibration function depending on synthetically crafted neighbors.

Formally, for a given x, we aim to produce natural adjacent samples, or a set of n neighbors $\{\tilde{x}_1, ..., \tilde{x}_n\}$, which slightly differ from x and are not part of the target model's training data, but are approximately equally likely to appear in the general distribution of textual data, and therefore offer a meaningful comparison. Given our set of neighbors, we calibrate the loss score of x under the target model by subtracting the average loss of its neighbors from it, resulting in a new decision rule:

$$A_{f_\theta}(x) = \mathbb{1}\left[\left(\mathcal{L}(f_\theta, x) - \sum_{i=1}^{n} \frac{\mathcal{L}(f_\theta, \tilde{x}_i)}{n}\right) < \gamma\right]. \qquad (3)$$

The interpretation of this decision rule is straightforward: Neighbors crafted through minimal changes that fully preserve the semantics and grammar of a given sample should in theory be interchangeable with the original sentence and therefore be assigned highly similar likelihoods under any textual probability distribution. Assuming that our neighbors were not present in the training data of the target model, we can therefore use the model score assigned to them as a proxy for what the original sample's loss should be if it was not present in the training data. The target sample's loss value being substantially lower than the neighbors' losses could therefore only be a result of overfitting and therefore the target sample being a training member. In this case, we expect the difference in Equation 3 to be below our threshold value γ

2.2 Obtaining Neighbour Samples

In the previous section, for a given text x, we assumed access to a set of adjacent samples $\{\tilde{x}_1, ..., \tilde{x}_n\}$. In this section we describe how those samples are generated. As it is highly important to consider neighbours that are approximately equally complex, it is important to mention that beyond the semantics of x, we should also preserve structure and syntax, and can therefore not simply consider standard textual style transfer or paraphrasing models. Instead, we opt for very simple word replacements that preserve semantics and fit the context of the original word well. For obtaining these replacements, we adopt the framework proposed by Zhou et al. (2019), who propose the use of transformer-based (Vaswani et al., 2017) masked language models (MLMs) such as BERT (Devlin et al., 2019) for lexical substitutions: Concretely, given a text $x := (w^{(1)}, ..., w^{(L)})$ consisting of L tokens, the probability $p_\theta(\tilde{w} = w^{(i)}|x)$ of token

$\tilde{w}$ as the word in position i can be obtained from the MLM's probability distribution $p(\mathcal{V}^{(i)}|x)$ over our token vocabulary $\mathcal{V}$ at position i. As we do not want to consider the influence of the probability of the original token on the token's suitability as a replacement when comparing it to other candidates, we normalize the probability over all probabilities except that of the original token. So, if $\hat{w}$ was the original token at position i, our suitability score for $\tilde{w}$ as a replacement is

$$p_{\text{swap}}(\hat{w}^{(i)}, \tilde{w}^{(i)}) = \frac{p_\theta(\tilde{w} = w^{(i)}|x)}{1 - p_\theta(\hat{w} = w^{(i)}|x)}. \quad (4)$$

In practice, simply masking the token which we want to replace will lead to our model completely neglecting the meaning of the original word when predicting alternative tokens and therefore potentially change the semantics of the original sentence – for instance, for the given sample "The movie was great", the probability distribution for the last token obtained from "The movie was [MASK]" might assign high scores to negative words such as "bad", which are clearly not semantically suitable replacements. To counteract this, Zhou et al. (2019) propose to keep the original token in the input text, but to add strong dropout to the input embedding layer at position i before feeding it into the transformer to obtain replacement candidates for $w^{(i)}$. We adopt this technique, and therefore obtain a procedure which allows us to obtain n suitable neighbors with m word replacements using merely an off-the-shelf model that does not require any adaptation to the target domain. The pseudocode is outlined in Algorithm 1.

Algorithm 1: Neighbourhood Generation

Input : Text $x = (w^{(1)}, ..., w^{(L)})$, n, m
Output: Neighbours $\{\tilde{x}_1, ..., \tilde{x}_n\}$ with m word replacements each

for $i \in \{1, \ldots, L\}$ **do**
  Get embeddings $(\phi(w^{(1)}), .., \phi(w^{(L)}))$.
  Add dropout: $\phi(w^{(i)}) = \text{drop}(\phi(w^{(i)}))$.
  Obtain $p(\mathcal{V}^{(i)}|x)$ from BERT.
  Compute $p_{\text{swap}}(w^{(i)}, \tilde{w}^{(i)}) \forall \tilde{w} \in \mathcal{V}$.

For all swaps $(w^{(i_1)}, \tilde{w}^{(i_1)})...(w^{(i_m)}, \tilde{w}^{(i_m)})$ with $i_k \neq i_l$ for $i \neq l$, compute joint suitability $\sum_{i=1}^{m} p_{\text{swap}}(w^{(i_1)}, \tilde{w}^{(i_1)})$ and return n highest

---

3 Experimental Setup

We evaluate the performance of our attack as well as reference-free and reference-based baseline attacks against large autoregressive models trained with the classical language modeling objective. Particularly, we use the base version of GPT-2 (Radford et al., 2019) as our target model.

3.1 Datasets

We perform experiments on three datasets, particularly news article summaries obtained from a subset of the AG News corpus[1] containing four news categories ("World", "Sports", "Business", "Science & Technology"), tweets from the Sentiment140 dataset (Go et al., 2009) and excerpts from wikipedia articles from Wikitext-103 (Merity et al., 2017). Both datasets are divided into two disjunct subsets of equal size: one of these subsets serves as training data for the target model and therefore consists of positive examples for the membership classification task. Subset two is not used for training, but its samples are used as negative examples for the classification task. The subsets contain 60,000, 150,000 and 100,000 samples for AG News, Twitter and Wikitext, respectively, leading to a total size of 120,000, 300,000 and 200,000 samples. For all corpora, we also keep an additional third subset that we can use to train reference models for reference-based attacks.

3.2 Baselines

To compare the performance of our attack, we consider various baselines: As the standard method for reference-free attacks, we choose the **LOSS Attack** proposed by Yeom et al. (2018), which classifies samples as training members or non-members based on whether their loss is above or below a certain threshold (see Equation 1). For reference-based attacks, we follow recent implementations (Mireshghallah et al., 2022a,b; Watson et al., 2022) and use reference data to train a single reference model of the same architecture as the target model. Subsequently, we measure whether the likelihood of a sample under the target model divided by its likelihood under the reference model crosses a certain threshold.

Training Data for Reference Models As discussed in previous sections, we would like to evaluate reference-based attacks with more realistic

---

[1] http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

assumptions about access to the training data distribution. Therefore, we use multiple reference models trained on different datasets: As our **Base Reference Model**, we consider the pretrained, but not fine-tuned version of GPT-2. Given the large pretraining corpus of this model, it should serve as a good estimator of the general complexity of textual samples and has also been successfully used for previous implementations of reference-based attacks (Mireshghallah et al., 2022b). Similar to our neighbourhood attack, this reference model does not require an attacker to have any additional data or knowledge about the training data distribution.

To train more powerful, but still realistic reference models, which we henceforth refer to as **Candidate Reference Models**, we use data that is in general similar to the target model's training data, but slightly deviates with regard to topics or artifacts that are the result of the data collection procedure. Concretely, we perform this experiment for both our AG News and Twitter corpora: For the former, we use article summaries from remaining news categories present in the AG News corpus ("U.S.", "Europe", "Music Feeds", "Health", "Software and Development", "Entertainment") as well as the NewsCatcher dataset[2] containing article summaries for eight categories that highly overlap with AG News ("Business", "Entertainment", "Health", "Nation", "Science", "Sports", "Technology", "World"). For Twitter, we use a depression detection dataset for mental health support from tweets [3] as well as tweet data annotated for offensive language [4]. As it was highly difficult to find data for reference models, it was not always possible to match the amount of training samples of the target model. The number of samples present in each dataset can be found in Table 1.

As our most powerful reference model, henceforth referred to as **Oracle Reference Model**, we use models trained on the same corpora, but different subsets as the target models. This setup assumes that an attacker has perfect knowledge about the training data distribution of the target model and high quality samples.

Dataset	#Samples
AG News (Other Categories)	60,000
NewsCatcher	60,000
AG News Oracle Data	60,000
Twitter Mental Health	20,000
Twitter Offensive Language	25,000
Twitter Oracle Data	150,000
Wikipedia Oracle Data	100,000

Table 1: Number of samples in the reference model training data. Target models for News, Twitter and Wikipedia were trained on 60,000, 150,000 and 100,000 samples, respectively.

3.3 Implementation Details

We obtain and fine-tune all pretrained models using the Huggingface transformers library (Wolf et al., 2020) and PyTorch (Paszke et al., 2019). As target models, we fine-tune the pretrained 117M parameter version of GPT-2, which originally has a validation perplexity of 56.8 and 200.3 on AG News and Twitter data, respectively, up to validation set perplexities of 30.0 and 84.7. In our initial implementation of our neighbourhood attack, we obtain the 100 most likely neighbour samples using one word replacement only from the pretrained 110M parameter version of BERT. We apply a dropout of $p = 0.7$ to the embedding of the token we want to replace. For evaluating LiRA baselines, we train each reference model on its respective training dataset over multiple epochs, and choose the best performing reference model w.r.t attack performance. Following Carlini et al. (2021a), we evaluate our attack's precision for predetermined low false positive rate values such as 1% or 0.01%. We implement this evaluation scheme by adjusting our threshold γ to meet this requirement and subsequently measure the attack's precision for the corresponding γ. All models have been deployed on single GeForce RTX 2080 and Tesla K40 GPUs.

4 Results

In this section, we report our main results and perform additional experiments investigating the impact of reference model performance on the success of reference-based attacks as well as several ablation studies. Following (Carlini et al., 2021a), we report attack performances in terms of their true positive rates (TPR) under very low false positive rates (FPR) by adjusting the threshold value γ. Concretely, we choose 1%, 0.1% and 0.01% as our target FPR values.

	News			Twitter			Wikipedia		
False Positive Rate	1%	0.1%	0.01%	1%	0.1%	0.01%	1%	0.1%	0.01%
Likelihood Ratio Attacks:									
Base Reference Model	4.24%	0.91%	0.16%	5.66%	0.98%	0.22%	1.21%	0.12%	0.01%
Candidate Reference Model 1	4.91%	0.95%	0.15%	6.49%	1.10%	0.24%			
Candidate Reference Model 2	4.76%	0.92%	0.15%	6.61%	1.19%	0.25%			
Oracle Reference Model*	18.90%	3.76%	0.16%	13.90%	1.59%	0.28%	11.70%	3.70%	0.12%
Reference-Free Attacks:									
LOSS Attack	3.50%	0.10%	0.01%	2.08%	0.11%	0.02%	1.06%	0.11%	0.01%
Neighbour Attack (Ours)	**8.29%**	**1.73%**	**0.29%**	**7.35%**	**1.43%**	**0.28%**	**2.32%**	**0.27%**	**0.10%**

Table 2: True positive rates of various attacks for low false positive rates of $1\%, 0.1\%$, and 0.01%. Candidate Reference Model 1 refers to reference models trained on data from other AG News categories and our Twitter mental health dataset, Candidate Reference Model 2 refers to reference models trained on NewsCatcher and the offensive tweet classification dataset. *As reference attacks trained on oracle datasets represent a rather unrealistic scenario with perfect assumptions, we compare our results with other baselines with more realistic assumptions when highlighting best results as bold.

4.1 Main Results

Our results can be found in Table 2 and 3, with the former showing our attack performance in terms of true positive rates under low false positive rates and the latter showing AUC values. As previously discovered, the LOSS attack tends to perform badly when evaluated for very low false positive rates (Carlini et al., 2021a; Watson et al., 2022). Likelihood Ratio Attacks can clearly outperform it, but we observe that their success is highly dependent on having access to suitable training data for reference models: Attacks using the base reference models and candidate models can not reach the performance of an attack using the oracle reference model by a large margin. Notably, they are also substantially outperformed by our Neighbour Attack, which can, particularly in low FPR ranges, even compete very well with or outperform Likelihood Ratio Attacks with an Oracle Reference Model, without relying on access to any additional data.

	News	Twitter	Wiki
LiRA:			
Base Reference Model	0.76	0.75	0.54
Candidate Reference 1	0.78	**0.81**	
Candidate Reference 2	0.75	0.77	
Oracle Reference*	0.94	0.89	0.90
Other Attacks:			
LOSS Attack	0.64	0.60	0.52
Neighbour Attack	**0.79**	0.77	**0.62**

Table 3: AUC values of various attacks.

4.2 Measuring the Dependence of Attack Success on Reference Model Quality

Motivated by the comparably poor performance of Likelihood Ratio Attacks with reference models trained on only slightly different datasets to the target training data, we aim to investigate the dependence of reference attack performances on the quality of reference models in a more controlled and systematic way. To do so, we train reference models on our oracle data over multiple epochs, and report the attack performance of Likelihood Ratio Attacks w.r.t to the reference models' validation perplexity (PPL) on a held out test set, which is in this case the set of non-training members of the target model. Intuitively, we would expect the attack performance to peak when the validation PPL of reference models is similar to that of the target model, as this way, the models capture a very similar distribution and therefore offer the best comparison to the attack model. In this setup, we are however particularly interested in the attack performance when the validation PPL does not exactly match that of the target model, given that attackers will not always be able to train perfectly performing reference models.

The results of this experiment can be found in Figure 2 for our News and Twitter dataset and in Figure 3 for Wikitext. As can be seen, the performance of reference-based attacks does indeed peak when reference models perform roughly the same as the target model. A further very interesting observation is that substantial increases in attack success only seem to emerge as the validation PPL of reference models comes very close to that of the target model and therefore only crosses the success
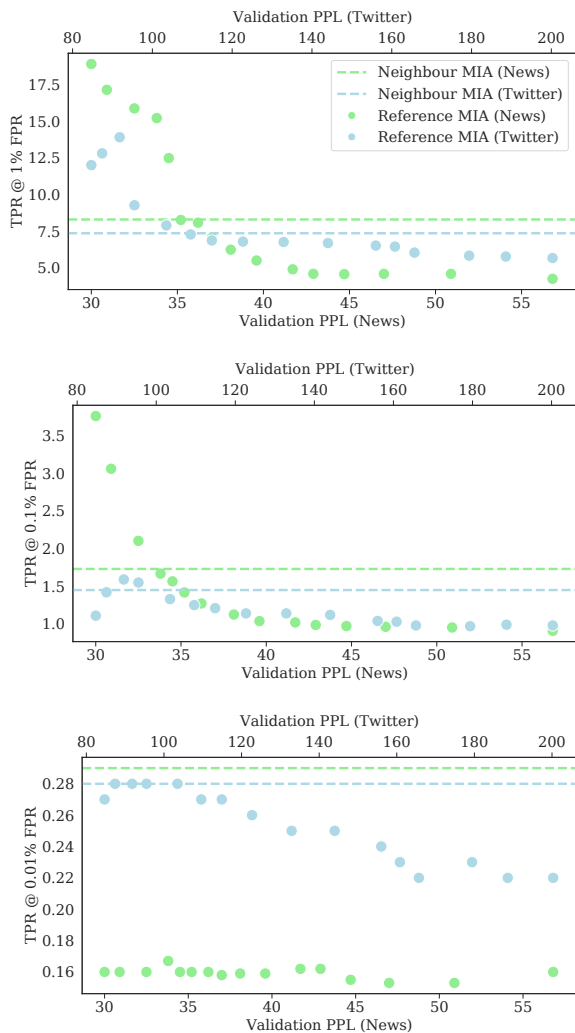
Figure 2: Attack Performance of reference attacks w.r.t validation PPL of reference models, compared to the performance of neighborhood attacks. The perplexities of the target models were 30.0 and 84.7 for AG News and Twitter, respectively

#Neighbours	5	10	25	50	100
News:					
1% FPR	2.98%	4.57%	6.65%	8.19%	8.29%
0.1% FPR	0.53%	0.79%	1.43%	1.50%	1.73%
0.01% FPR	0.05%	0.07%	0.18%	0.23%	0.29%
Twitter:					
1% FPR	3.93%	4.88%	6.21%	6.63%	7.35%
0.1% FPR	0.57%	0.62%	1.01%	1.34%	1.43%
0.01% FPR	0.05%	0.07%	0.10%	0.23%	0.28%
Wikipedia:					
1% FPR	1.57%	1.81%	2.02%	2.17%	2.32%
0.1% FPR	0.16%	0.21%	0.23%	0.26%	0.27%
0.01% FPR	0.05%	0.08%	0.09%	0.10%	0.10%

Table 4: Attack performance w.r.t the number of neighbours against which we compare the target sample

erated neighbours as determined by BERT. In the following, we measure how varying this number affects the attack performance. While intuitively, a higher number of neighbours might offer a more robust comparison, it is also plausible that selecting a lower number of most likely neighbours under BERT will lead to neighbours of higher quality and therefore a more meaningful comparison of loss values. Our results in Table 4 show a clear trend towards the former hypothesis: The number of neighbours does in general have a strong influence on the performance of neighbourhood attacks and higher numbers of neighbours produce better results.

Number of Word Replacements Besides the number of generated neighbours, we study how the number of replaced words affects the performance of our attack. While we reported results for the replacement of a single word in our main results in Table 2, there are also reasons to expect that a higher number of replacements leads to better attack performance: While keeping neighbours as similar to the original samples as possible ensures that their probability in the general distribution of textual data remains as close as possible, one could also expect that too few changes lead the target model to assign the original sample and its neighbours almost exactly the same score, and therefore make it hard to observe high differences in loss scores for training members. Our results of generating 100 neighbours with multiple word replacements are reported in Table 5. We find that replacing only one word clearly outperforms multiple replacements. Beyond this, we do not find highly meaningful differences between two and three word replacements.

rate of neighbourhood attacks when the reference model's performance is almost the same as that of the target model. This further illustrates the fragility of reference-based attacks with respect to the choice of the reference model.

4.3 Ablation Studies

Having extensively studied the impact of different reference model training setups for the Likelihood Ratio Attack, we now aim to explore the effect of various components of our proposed neighbourhood attack.

Number of Generated Neighbours For our main results in Table 2, we report the performance of neighbour attacks for the 100 most likely gen-

#Word Replacements	1	2	3
News:			
1% FPR	8.29%	4.09%	4.18%
0.1% FPR	1.73%	0.85%	0.94%
0.01% FPR	0.29%	0.23%	0.21%
Twitter:			
1% FPR	7.35%	4.86%	4.37%
0.1% FPR	1.43%	0.74%	0.72%
0.01% FPR	0.28%	0.14%	0.11%
Wikipedia:			
1% FPR	2.32%	1.76%	1.44%
0.1% FPR	0.27%	0.23%	0.17%
0.01% FPR	0.10%	0.07%	0.03%

Table 5: Attack performance w.r.t the number of words that are replaced when generating neighbours

	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = \infty$
TPR @ 1% FPR	1.29%	1.52%	8.29%
TPR @ 0.1% FPR	0.09%	0.13%	1.73%
TPR @ 0.01% FPR	0.01%	0.01%	0.29%

Table 6: Performance of neighbourhood attacks against models trained with DP-SGD

5 Defending against Neighbourhood Attacks

Due to the privacy risks that emerge from the possibility of membership inference and data extraction attacks, the research community is actively working on defenses to protect models. Beyond approaches such as confidence score perturbation (Jia et al., 2019) and specific regularization techniques (Mireshghallah et al., 2021; Chen et al., 2022) showing good empirical performance, differentially private model training is one of the most well known defense techniques offering mathematical privacy guarantees: DP-SGD (Song et al., 2013; Bassily et al., 2014; Abadi et al., 2016), which uses differential privacy (Dwork et al., 2006) to bound the influence that a single training sample can have on the resulting model and has been shown to successfully protect models against membership inference attacks (Carlini et al., 2021a) and has recently also successfully been applied to training language models (Yu et al., 2022; Li et al., 2022; Mireshghallah et al.). To test the effectiveness of differential privacy as a defense against neighbourhood attacks, we follow Li et al. (2022) and train our target model GPT-2 in a differentially private manner on AG News, where our attack performed the best. The results can be seen in Table 6 and clearly demonstrate the effectiveness of DP-SGD. Even for comparably high epsilon values such as ten, the performance of the neighbourhood attack is substantially worse compared to the non-private model and is almost akin to random guessing for low FPR values.

6 Related Work

MIAs have first been proposed by Shokri et al. (2016) and continue to remain a topic of interest for the machine learning community. While many attacks, such as ours, assume to only have access to model confidence or loss scores (Yeom et al., 2018; Sablayrolles et al., 2019; Jayaraman et al., 2020; Watson et al., 2022), others exploit additional information such as model parameters (Leino and Fredrikson, 2020) or training loss trajectories (Liu et al., 2022). Finally, some researchers have also attempted to perform membership inference attacks given only hard labels without confidence scores (Li and Zhang, 2021; Choquette-Choo et al., 2021). Notably, the attack proposed by Choquette-Choo et al. (2021) is probably closest to our work as it tries to obtain information about a sample's membership by flipping its predicted labels through small data augmentations such as rotations to image data. To the best of our knowledge, we are the first to apply data augmentations of this kind for text-based attacks.

Membership Inference Attacks in NLP Specifically in NLP, membership inference attacks are an important component of language model extraction attacks (Carlini et al., 2021b; Mireshghallah et al., 2022b). Further studies of interest include work by Hisamoto et al. (2020), which studies membership inference attacks in machine translation, as well as work by Mireshghallah et al. (2022a), which investigates Likelihood Ratio Attacks for masked language models. Specifically for language models, a large body of work also studies the related phenomenon of memorization (Kandpal et al., 2022; Carlini et al., 2022b,a; Zhang et al., 2021), which enables membership inference and data extraction attacks in the first place.

Machine-Generated Text Detection Due to the increasing use of tools like ChatGPT as writing assistants, the field of machine-generated text detection has become of high interest within the research community and is being studied extensively

(Chakraborty et al., 2023; Krishna et al., 2023; Mitchell et al., 2023; Mireshghallah et al., 2023). Notably, Mitchell et al. (2023) propose DetectGPT, which works similarly to our attack as it compares the likelihood of a given sample under the target model to the likelihood of perturbed samples and hypothesizes that the likelihood of perturbations is smaller than that of texts the model has generated itself.

7 Conclusion and Future Work

In this paper, we have made two key contributions: First, we thoroughly investigated the assumption of access to in-domain data for reference-based membership inference attacks: In our experiments, we have found that likelihood ratio attacks, the most common form of reference-based attacks, are highly fragile to the quality of their reference models and therefore require attackers to have access to high-quality training data for those. Given that specifically in privacy-sensitive settings where publicly available data is scarce, this is not always a realistic assumption, we proposed that the design of reference-free attacks would simulate the behavior of attackers more accurately. Thus, we introduced neighborhood attacks, which calibrate the loss scores of a target samples using loss scores of plausible neighboring textual samples generated through word replacements, and therefore eliminate the need for reference trained on in-domain data. We have found that under realistic assumptions about an attacker's access to training data, our attack consistently outperforms reference-based attacks. Furthermore, when an attacker has perfect knowledge about the training data, our attack still shows competitive performance with reference-based attacks. We hereby further demonstrated the privacy risks associated with the deployment of language models and therefore the need for effective defense mechanisms. Future work could extend our attack to other modalities, such as visual or audio data, or explore our attack to improve extraction attacks against language models.

Limitations

The proposed attack is specific to textual data While many membership inference attacks are universally applicable to all modalities as they mainly rely on loss values obtained from models, our proposed method for generating neighbours is specific to textual data. While standard augmentations such as rotations could be used to apply our method for visual data, this is not straightforward such as the transfer of other attacks to different modalities.

Implementation of baseline attacks As the performance of membership inference attacks depend on the training procedure of the attacked model as well as its degree of overfitting, it is not possible to simply compare attack performance metrics from other papers to ours. Instead, we had to reimplement existing attacks to compare them to our approach. While we followed the authors' descriptions in their papers as closely as possible, we cannot guarantee that their attacks were perfectly implemented and the comparison to our method is therefore 100% fair.

Ethical Considerations

Membership inference attacks can be used by malicious actors to compromise the privacy of individuals whose data has been used to train models. However, studying and expanding our knowledge of such attacks is crucial in order to build a better understanding for threat models and to build better defense mechanisms that take into account the tools available to malicious actors. Due to the importance of this aspect, we have extensively highlighted existing work studying how to defend against MIAs in Section 6. As we are aware of the potential risks that arise from membership inference attacks, we will not freely publicize our code, but instead give access for research projects upon request.

With regards to the data we used, we do not see any issues as all datasets are publicly available and have been used for a long time in NLP research or data science competitons.

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*, CCS '16, page 308–318, New York, NY, USA. Association for Computing Machinery.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. 2014. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, A. Terzis, and Florian Tramèr. 2021a. Membership inference attacks from first principles. *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022a. Quantifying memorization across neural language models.

Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramer. 2022b. The privacy onion effect: Memorization is relative. In *Advances in Neural Information Processing Systems*.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021b. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023. On the possibilities of ai-generated text detection.

Dingfan Chen, Ning Yu, and Mario Fritz. 2022. Relaxloss: Defending membership inference attacks without losing utility. In *International Conference on Learning Representations*.

Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. 2021. Label-only membership inference attacks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1964–1974. PMLR.

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.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg. Springer Berlin Heidelberg.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

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.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Bargav Jayaraman, Lingxiao Wang, David E. Evans, and Quanquan Gu. 2020. Revisiting membership inference under realistic assumptions. *Proceedings on Privacy Enhancing Technologies*, 2021:348 – 368.

Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. 2019. Memguard: Defending against black-box membership inference attacks via adversarial examples. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense.

Klas Leino and Matt Fredrikson. 2020. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *Proceedings of the 29th USENIX Conference on Security Symposium*, SEC'20, USA. USENIX Association.

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*.

Zheng Li and Yang Zhang. 2021. Membership leakage in label-only exposures. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, page 880–895, New York, NY, USA. Association for Computing Machinery.

Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. 2022. Membership Inference Attacks by Exploiting Loss Trajectory. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 2085–2098. ACM.

Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A. Gunter, and Kai Chen. 2018. Understanding membership inferences on well-generalized learning models. *ArXiv*, abs/1802.04889.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Fatemehsadat Mireshghallah, Arturs Backurs, Huseyin A Inan, Lukas Wutschitz, and Janardhan Kulkarni. Differentially private model compression. In *Advances in Neural Information Processing Systems*.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022a. Quantifying privacy risks of masked language models using membership inference attacks.

Fatemehsadat Mireshghallah, Huseyin Inan, Marcello Hasegawa, Victor Rühle, Taylor Berg-Kirkpatrick, and Robert Sim. 2021. Privacy regularization: Joint privacy-utility optimization in LanguageModels. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3799–3807, Online. Association for Computational Linguistics.

Fatemehsadat Mireshghallah, Justus Mattern, Sicun Gao, Reza Shokri, and Taylor Berg-Kirkpatrick. 2023. Smaller language models are better black-box machine-generated text detectors. *arXiv preprint arXiv:2305.09859*.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. 2022b. Memorization in nlp fine-tuning methods.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature.

Sasi Kumar Murakonda and R. Shokri. 2020. Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. *ArXiv*, abs/2007.09339.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. 2019. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*.

R. Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2016. Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, page 377–390, New York, NY, USA. Association for Computing Machinery.

Shuang Song, Kamalika Chaudhuri, and Anand D. Sarwate. 2013. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248.

Shuang Song and David Marn. 2020. Introducing a new privacy testing library in tensorflow.

Florian Tramèr, Gautam Kamath, and Nicholas Carlini. 2022. Considerations for differentially private learning with large-scale public pretraining.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. 2022. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. 2022. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, page 3093–3106, New York, NY, USA. Association for Computing Machinery.

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.
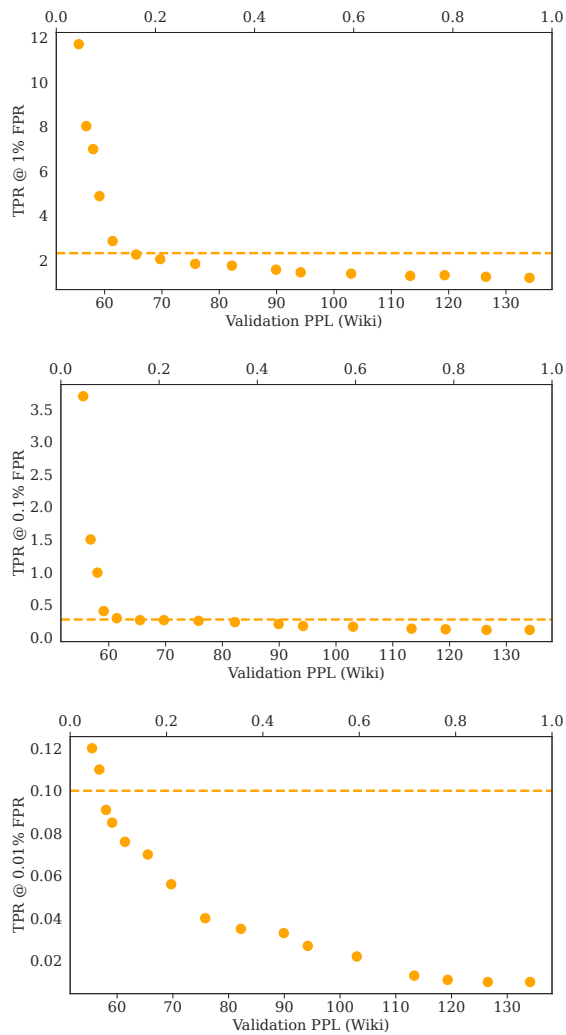
Figure 3: Attack Performance of reference attacks w.r.t validation PPL of reference models, compared to the performance of neighborhood attacks. The perplexity of the target model was 55.6 for Wikipedia

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2022. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2021. Counterfactual memorization in neural language models.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.

A   For every submission:

☑ A1. Did you describe the limitations of your work?
"Limitations" Section

☑ A2. Did you discuss any potential risks of your work?
Introduction (Section 1), Ethical Considerations Section and Limitation Section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
See abstract and Introduction (Section 1)

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B   ☑ Did you use or create scientific artifacts?

Used BERT and GPT2 from Huggingface, as well as benchmark datasets for news headlines and tweets (see section 3, Experiments). We also wrote code, which we refer to in the implementation details in section 3.3

☑ B1. Did you cite the creators of artifacts you used?
see section 3 (Experiments), particularly subsections Implementation Details and Datasets (3.3 and 3.1)

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We provide links referring to the webpages of all datasets used in our paper. The license and terms of use can immediately be found on this page.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section "Ethical Considerations": We confirm that all datasets are established benchmarks in research

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
As discussed in the Section "Ethical Considerations", the data are established benchmarks and widely used. Therefore, these aspects have been covered in prior work - additionally, one of our datasets (Twitter offensive language classification) does contain offensive contents, as it is the purpose of the dataset. We argue that for the intended use cases of hate speech detection, such datasets benefit the positive impact of research.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
As the data we use is very simple (it contains news headlines and tweets), such information is not very extensive and out of scope for our paper, as it is not highly relevant for membership inference attacks. Such analysis can also be found in the cited papers and websites presenting the benchmarks

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader

to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3 (Experiments), particularly 3.1 and 3.2

C ☑ **Did you run computational experiments?**

Section 3 (Experiments)

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3 (Experiments), particularly 3.3

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3 (Experiments), particularly 3.3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Results Section (Section 4)

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3 (Experiments), particularly 3.3

D ☒ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.