

## A Background

### A.1 Private LLM Adaptations

Differentially Private Stochastic Gradient Descent (DP-SGD) [2] is a widely used method for incorporating DP into deep learning. However, while applied to NLP tasks, DP-SGD can exhibit several limitations, particularly in model utility, increased memory usage, or slower convergence during training. These limitations motivate the exploration of alternative DP adaptation techniques.

**Full DP Fine-Tuning.** One approach to differentially private (DP) adaptation is to fine-tune the entire model using the DPSGD algorithm [2, 30, 54]. This method updates all model parameters while ensuring that each gradient step satisfies DP guarantees through gradient clipping and noise addition. Full-model DP fine-tuning provides high adaptability and task-specific performance. However, it is computationally expensive and memory-intensive, especially for large language models (LLMs), due to the need to compute, clip, and perturb gradients for all layers [30].

**DP Head Fine-Tuning.** An alternative strategy is to fine-tune only the final layer (often called the classification or task-specific “head”) of the model using DP-SGD. This significantly reduces the number of trainable parameters, leading to lower memory usage and faster training. Despite its simplicity, DP Head Fine-Tuning can still achieve competitive performance on certain tasks while providing formal privacy guarantees. However, its adaptability is limited, particularly when deeper model layers need task-specific adjustments.

**DP low-rank adaptation (LoRA).** LoRA [21] is an efficient technique for adapting LLMs that introduces low-rank matrices into each layer of a frozen pretrained model. Instead of updating the full weight matrix  $W \in \mathbb{R}^{d \times k}$ , LoRA learns a low-rank approximation  $\Delta W = AB$ , where  $A \in \mathbb{R}^{d \times r}$ ,  $B \in \mathbb{R}^{r \times k}$ , and  $r \ll \min(d, k)$ . The adapted weights become  $W' = W + AB$ , with only  $A$  and  $B$  being trainable. DP LoRA [53] extends this approach by applying DPSGD to the low-rank parameters. This ensures that the adaptation remains privacy-preserving, making LoRA suitable for sensitive-data applications with formal DP guarantees.

**DP Prompting.** Introducing a small set of additional parameters, typically under 1% of the LLMs total parameters, DP Prompting applies these only within the model’s input space. These parameters may be added at the level of token embeddings (soft prompts [31, 32]) or to all (attention) layers of the LLM (prefix-tuning [26, 29]). Duan et al. [13] proposed *PromptDPSGD*, which adapts the DPSGD algorithm [2] for use with soft prompts.

### A.2 MIAs

The following section provides a more detailed description of MIAs used in our benchmark.

**Min-K%.** Min-K% [42] is a recently proposed black-box MIA for large language models. The intuition is that an unseen sample is likely to have low-probability tokens. The MIA score is defined as

$$\text{Min-K\%}(x) = \frac{1}{|S|} \sum_{x_i \in S} \log p(x_i | x_1, \dots, x_{i-1}), \quad (3)$$

where  $S$  is the set of  $K\%$  tokens with the smallest loss.

**Reference.** This approach [7] uses a reference model to calibrate the MI score as follows

$$\text{Ref}(x) = \frac{\mathcal{L}(x|\theta)}{\mathcal{L}(x|\theta_{\text{ref}})}, \quad (4)$$

where  $\mathcal{L}(x|\theta)$  indicates the loss of the target sample  $x$  on the model  $\theta$ .  $\theta_{\text{ref}}$  represents the reference model used.

**Robust Membership inference attack (RMIA).** RMIA outperforms previous methods by optimizing the null hypothesis and using a reference model along with population data, requiring only one reference (*shadow*) model at a time, unlike previous methods [8] which required hundreds. RMIA has two hyperparameters, a threshold  $\gamma$  and a scaling factor  $\alpha$ . The adapted RMIA score (Equation (5)) calculation for LLMs for text generation is based on comparing loss values rather than output probabilities. For this reason, we have to, instead of comparing prediction probabilities or logits, compare the loss of the target data point against the loss of reference models on population data

(Equation (6)) and flip to a minority voting approach, where the decision is based on how much lower the loss of the target data is compared to the population data.

$$\text{Score}_{\text{MIA}}(x; \theta) = \Pr_{z \sim \pi} (\text{LR}_{\theta}(x, z) \geq \gamma) \quad (5)$$

$$\text{LR}_{\theta}(x, z) = \mathcal{L}(\theta|x) - \mathcal{L}(\theta|z) \quad (6)$$

### A.3 Canary Exposure and Data Extraction Attacks

Following Carlini et al. [6], Tramèr et al. [47], let  $\mathcal{U}$  be the universe of candidate samples and let  $\hat{Z}$  be the attacker’s ranking of  $\mathcal{U}$  by model-assigned likelihood. For a target  $z \in \mathcal{U}$ ,

$$\text{exposure}(z, \hat{Z}) := \log_2 |\mathcal{U}| - \log_2 (\text{rank}(z; \hat{Z})). \quad (7)$$

This metric ranges from 0 (least likely) to  $\log_2 |\mathcal{U}|$  (most likely). To compute it efficiently when  $|\mathcal{U}|$  is large, one can use: (1) **sampling**, which estimates exposure on a random subset of  $\mathcal{U}$ , or (2) **distribution modeling**, which approximates the distribution of model scores (e.g. via a skewed normal) to interpolate ranks. The expected exposure of an unmemorized canary is  $\frac{1}{\ln 2} \approx 1.44$  [22]. Complementing exposure-based metrics, Carlini et al. [9] introduce a contextual extraction framework to assess memorization and data extraction attacks. Let  $f$  be a generative model and  $s$  a secret suffix. We say  $s$  is *extractable with  $k$  tokens of context* if there exists a prefix  $p$  of length  $k$  such that, under greedy decoding,

$$f(p) = [p \parallel s].$$

When  $s$  is long and random, its successful extraction indicates memorization. One can vary  $k$  to characterize how much context the model needs before regurgitating  $s$  verbatim.

## B Additional Details on the Setup

### B.1 Datasets

For the IID datasets, we focus on the following Pile subsets: BookCorpus2, consisting of publicly available books, GitHub, a set of open-source code repositories, and Enron Emails [24], various emails. The OOD datasets we choose for our experiments are: SAMSum [19], an English-language dialogue summarization dataset, and GermanWiki [1], a large set of German Wikipedia entries. These OOD datasets were selected because of their different degrees of variation from the original distribution of the Pile dataset. Although SAMSum shares the same language (English), its general dialogue format, followed by the dialogue summary, is not present in the pretraining set. GermanWiki, on the other hand, presents wide syntactic and lexical variation from the pretraining dataset.

### B.2 Adaptations

We focus on four types of adaptations: Prefix Tuning, LoRA, Full Fine-Tune, and Head Fine-Tune. We train all the models using Adam with the privatization gradient method of DPSGD [2]. For the Adam optimizer, we use the default HuggingFace hyperparameters except for the learning rate. For Prefix Tuning, we fix a prefix length of 64, while for LoRA, a rank  $r = 8$  and  $\alpha = 16$ . For DP-SGD, following existing work [30], we set the gradient clipping value to 0.1. Moreover, in all settings, we consider sentence-level DP, meaning that we concatenate all strings in the dataset and split them into 256 token chunks, corresponding to sentence-level privacy.

### B.3 Hyperparameters

For each task, model, and privacy budget, we performed a hyperparameter optimization using a random search strategy. Specifically, we explored the following ranges:

- Learning Rate:  $1 \times 10^{-6}, 3 \times 10^{-6}, 1 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}, 8 \times 10^{-5}, 1 \times 10^{-4}, 3 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}$ ;
- Number of training epochs: 1, 2, 3, 5, 10, 15, 16, 20, 30, 32;

658 • Batch size: 4, 8, 16, 32, 64;

659 Our objective during hyperparameter search is to ensure comparable evaluation perplexities, specifically targeting similar validation loss values after adaptation training across different methods for  
660 specific datasets.  
661

## 662 B.4 MIA

663 The adopted offline mode (see Algorithm 1) shrinks from the need to retrain reference models per  
664 query, thus relying on pretrained LLMs, which are computationally expensive to train. For most  
665 experiments, we used just one reference model ( $k = 1$ ), thus demonstrating the power of RMIA  
666 attack and highlighting data leakage, especially from pretrained data. For an ablation on the RMIA  
667 hyperparameters choice, see Figure 13 in Appendix H.

---

**Algorithm 1** MIA score calculation with offline RMIA [56] adapted to LLMs.

---

**Input:**  $k$  reference models  $\Theta$ , target sample  $x$ , threshold  $\gamma$ , scaling factor  $\alpha$ , population dataset  $\pi$ ,

**Output:**  $\text{Score}_{\text{MIA}}(x; \theta)$

```

1: Randomly choose a subset  $Z$  from the population dataset
2:  $C \leftarrow 0$ 
3:  $\mathcal{L}(x)_{\text{OUT}} \leftarrow \frac{1}{k} \sum_{\theta' \in \Theta} \mathcal{L}(x|\theta')$ 
4:  $\mathcal{L}(x) \leftarrow \frac{1}{2} ((1 + \alpha)\mathcal{L}(x)_{\text{OUT}} + (1 - \alpha))$ 
5:  $\text{Ratio}_x \leftarrow \frac{\mathcal{L}(x|\theta)}{\mathcal{L}(x)}$ 
6: for each sample  $z$  in  $Z$  do
7:    $\mathcal{L}(z) \leftarrow \frac{1}{k} \sum_{\theta' \in \Theta} \mathcal{L}(z|\theta')$ 
8:    $\text{Ratio}_z \leftarrow \frac{\mathcal{L}(z|\theta)}{\mathcal{L}(z)}$ 
9:   if  $\text{Ratio}_x / \text{Ratio}_z < \gamma$  then
10:      $C \leftarrow C + 1$ 
11:   end if
12: end for
13: return  $\text{Score}_{\text{MIA}}(x; \theta) \leftarrow \frac{C}{|Z|}$ 

```

---

## 668 B.5 Canary Exposure

669 We add an adversarial prefix to  $p = 1\%$  of the adaptation data. If not specified otherwise, we set the  
670 number of canary tokens to  $k = 10$  and the canary prefix length  $l = 10$ . To measure exposure, we  
671 generate 256 new canary prefixes from the same canary type and prepend them to the target sample  
672  $x$  whose exposure we want to measure. The resulting 256 samples can be considered as a form of  
673 non-members. On expectation, all canary prefixes are equally (un)likely. However, if the model is  
674 more confident about the one prefix it saw during adaptation than it is about the other 256 prefixes, it  
675 means that the model must have memorized this prefix and that it was part of the adaptation data.  
676 Given that there are two ways of approximating exposure (sampling and distribution modeling) as  
677 discussed in Section 2, we assess both of them to find whether one approach is more suitable. This  
678 ablation in Figure 12 Appendix F shows that the two approximations perform similarly when using  
679 256 non-member canaries. In our experiments, we evaluated using *sampling* as an approximation  
680 since it is computationally cheaper.

681 **Canary Types.** The *random* canary prefix is the simplest type of canary prefix, and it is composed  
682 of completely random tokens sampled uniformly from the token universe  $T$ . The *common* and *rare*  
683 prefixes comprise the most and least frequently occurring tokens, respectively, excluding special  
684 tokens *e.g.*, padding and end-of-string tokens. We count the total number of token occurrences in the  
685 adaptation dataset to measure the frequencies. Then, we choose the top  $k$  tokens from a list sorted in  
686 ascending or descending order for *rare* and *common*, respectively. Note that, for both *common* and  
687 *rare*, each adaptation dataset naturally has its own set of distinct prefix tokens. We also select the  
688 *random* tokens independently over each adaptation dataset for symmetry. The *invisible* canary prefix  
689 utilizes imperceptible Unicode symbols or space-like tokens, such as zero-width spaces or zero-width  
690 non-joiners, which are nearly undetectable by humans, thus incorporating the design approach known

691 from other adversarial attacks [5]. Compared to the other canary types, the set of tokens is the same  
 692 for each dataset. Again, we randomly sample  $k$  imperceptible symbols to prepend as a canary prefix.

693 **Canary Adaptation Set Generation.** Algorithm 2 describes the procedure to construct the  
 694 adaptation dataset with canary prefixes. Note that  $\text{concat}(a,b)$  concatenates two strings, and the  
 695 tokens universe  $T$  represents the set of all the tokens accepted by the LLM. We prepend the canaries  
 696 to a small fraction  $p$  of the adaptation dataset prior to performing the adaptation. To each selected  
 697 sample, we add  $l$  many tokens, randomly drawn with replacement from the respective  $k$  canaries in  
 698 the canary prefix sets. We do not combine tokens from our four different types of canary prefixes and  
 699 consider each separately.

---

**Algorithm 2** Adding canary prefixes to the adaptation dataset.

---

**Input:**  $D$  adaptation dataset,  $t$  canary prefix type,  $l$  canary prefix length,  $k$  number of selected canaries,  $p$  canary prefix probability,  $T$  token universe.

**Output:**  $\tilde{D}$  modified adaptation dataset

```

1: if  $t = \text{"random"}$  then
2:    $C \leftarrow$  Randomly sample  $k$  tokens from  $T$ 
3: else if  $t = \text{"rare"}$  then
4:    $C \leftarrow$  Select the  $k$  least frequent tokens from  $D$ 
5: else if  $t = \text{"common"}$  then
6:    $C \leftarrow$  Select the  $k$  most frequent tokens from  $D$ 
7: else if  $t = \text{"invisible"}$  then
8:    $C \leftarrow$  Randomly sample  $k$  invisible tokens from  $T$ 
9: end if
10:  $D_0, D_1 \leftarrow$  Randomly split  $D$  in two datasets s.t. each
    sample is with probability  $p$  in  $D_1$ 
11:  $\tilde{D}_1 \leftarrow \{\}$ 
12: for each sample  $x \in D_1$  do
13:    $y \leftarrow$  Sample with replacement  $l$  tokens from  $C$ 
14:    $\tilde{D}_1 \leftarrow \tilde{D}_1 \cup \{\text{concat}(y, x)\}$ 
15: end for
16: return  $D_0 \cup \tilde{D}_1$ 

```

---

## 700 B.6 Extractable Memorization

701 Another privacy concern shown in prior work [9] is the memorization of samples during pretraining  
 702 of an LLM. We analyze how adaptations can reduce the effect of memorizing pretraining data. The  
 703 definition of a memorized sample follows  $k$ -extractability [9]. Here, we have a prompt  $p$  of length  
 704  $k$  and a suffix  $s$ . If the generation of a model given prompt  $p$  generates exactly  $s$ , the sequence  
 705 consisting of  $p$  and  $s$  concatenated is memorized.

706 We report the number of identified memorized samples for each Pile subset and Pythia 1B in Table 27  
 707 (Appendix G). Furthermore, we also rely on samples from the Pile reported as memorized in Pythia  
 708 2.8B by prior work [11]. This set of memorized samples consists of 505 sequences, and we refer to it  
 709 as Mem Pile.

## 710 B.7 Computational setup

711 We conduct most of our experiments on a single 40GB NVIDIA A100 GPU. However, for larger  
 712 models, we utilized a single NVIDIA A100 80GB Tensor Core GPU. The training time of the  
 713 adaptations varies depending on the applied adaptation method, the model size, the hyperparameters,  
 714 and whether DP is applied.

## 715 C Additional Experimental Results

### 716 C.1 MIAs

717 Table 3 and Table 4 present the MIA performance on OOD and IID datasets for the Pythia 1B model.  
 718 We repeat these experiments with other models from the Pythia [3] and GPT Neo [4] families to  
 719 broaden our study. Our findings include results for Pythia 1.4B (Table 5-Table 6), Pythia 410M

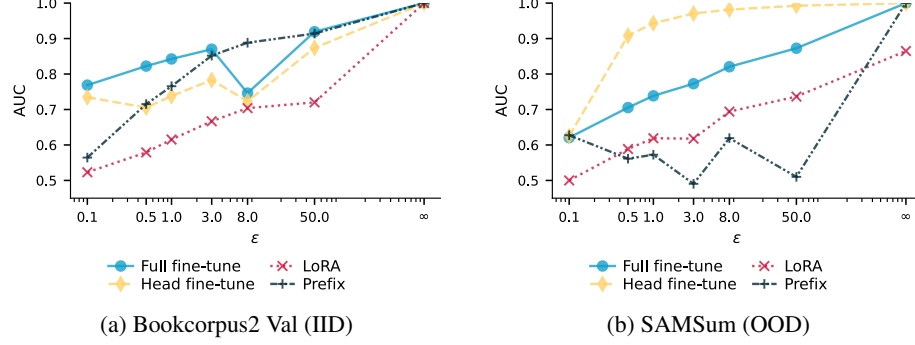


Figure 6: **The protection against MIA even for out-of-distribution (OOD) data requires tight privacy with  $\varepsilon < 0.1$  for all the adaptations.** The x-axis represents the privacy budget with a log scale, and the y-axis is the AUC score. The evaluation was done on Pythia 1B for  $\varepsilon = \{0.1, 0.5, 1, 3, 8, 50\}$ .

(Table 7-Table 8), Pythia 160M (Table 9-Table 10), Pythia 70M (Table 11-Table 12), GPT Neo 1.3B (Table 13-Table 14), and GPT Neo 125M (Table 15-Table 16). Our results indicate a privacy risk while adapting LLMs, and an attacker has advantages such as architectural knowledge, direct data access, and an exact understanding of the data split, thus allowing for a powerful attack vector. LoRA and Prefix are consistently less prone to MIA among most of the evaluated models and datasets than Full Fine-Tuning and Head-Fine-Tuning.

Overall, we observe a similar pattern between Pythia 1B, and the other evaluated models. For instance, for Pythia 410M (Table 7- Table 8), looking at *RMIA (shadow)* using  $\varepsilon = 8$ , we observe that the average AUC is 0.83, while for IID it is 0.9. Similarly, for Pythia 160M (Table 9- Table 10), the average AUC is 0.71 for OOD and 0.81 for IID data. These results follow our general trend that IID data taken from the pretraining validation set leaks just as much as data that directly overlaps, thus suggesting distributional closeness as the determining factor of privacy risk. Occasionally, we observe an anomaly, like the AUC for SAMSum in Table 5 being better under a privacy regime ( $\varepsilon = 8$ ) than without privacy protection. This behavior is a consequence of the fact that the loss is higher for the  $\varepsilon = \infty$  than for  $\varepsilon = 8$ . We prioritize having similar loss values across different adaptations for the given dataset and privacy budget. However, in some cases, the span of hyperparameters is too large to ensure that we have a similar loss across different  $\varepsilon$  values.

Going further, we also evaluate protection under varying privacy budgets, specifically  $\varepsilon \in \{0.1, 0.5, 1, 3, 8, 50\}$ . As illustrated in Figure 6, effective defense against privacy attacks, such as MIA, even for OOD data, requires a tight privacy bound of  $\varepsilon \leq 0.1$  for all adaptation strategies evaluated.

Table 3: **Membership Inference for OOD Adaptations.** We audit only the adaptations and assume the same pretrained LLM is used for all adaptations. We present the AUC scores obtained with reference, and Min-K% MIAs for the Pythia 1B model adapted on different datasets with  $\varepsilon \in \{0.1, 8, \infty\}$ .

MIA	Adaptation	Dataset			GermanWiki			Average		
		$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$
RMIA (shadow)	Prefix Tuning	1.00	0.62	0.63	1.00	0.64	0.61	1.00	0.63	0.62
	LoRA	0.86	0.69	0.50	1.00	0.59	0.66	0.93	0.64	0.58
	Full Fine-Tune	1.00	0.82	0.62	1.00	0.71	0.55	1.00	0.77	0.59
	Head Fine-Tune	1.00	0.98	0.62	1.00	0.76	0.70	1.00	0.87	0.66
	Average	0.97	0.78	0.59	1.00	0.67	0.63	0.98	0.73	0.61
RMIA (Pythia 1B)	Prefix Tuning	0.94	0.51	0.51	0.91	0.50	0.50	0.92	0.50	0.51
	LoRA	0.51	0.51	0.51	0.81	0.51	0.51	0.66	0.51	0.51
	Full Fine-Tune	0.94	0.51	0.51	0.98	0.51	0.51	0.96	0.51	0.51
	Head Fine-Tune	0.96	0.52	0.51	0.97	0.51	0.50	0.97	0.52	0.50
	Average	0.84	0.51	0.51	0.92	0.51	0.50	0.88	0.51	0.51
Reference (Pythia 1B)	Prefix Tuning	0.93	0.50	0.51	0.92	0.50	0.50	0.92	0.50	0.50
	LoRA	0.51	0.51	0.51	0.82	0.51	0.51	0.66	0.51	0.51
	Full Fine-Tune	0.94	0.51	0.51	0.99	0.51	0.50	0.96	0.51	0.51
	Head Fine-Tune	0.97	0.52	0.51	0.98	0.51	0.50	0.97	0.51	0.50
	Average	0.84	0.51	0.51	0.93	0.51	0.50	0.88	0.51	0.51
Min-K%	Prefix Tuning	0.84	0.51	0.51	0.71	0.50	0.50	0.78	0.50	0.50
	LoRA	0.51	0.51	0.50	0.61	0.51	0.51	0.56	0.51	0.51
	Full Fine-Tune	0.83	0.51	0.50	0.88	0.51	0.50	0.86	0.51	0.50
	Head Fine-Tune	0.92	0.51	0.50	0.87	0.51	0.51	0.89	0.51	0.50
	Average	0.77	0.51	0.50	0.77	0.50	0.51	0.77	0.51	0.50

Table 4: **Membership Inference for in-distribution (IID) Adaptations.** We use the same setup as in Table 3.

MIA	Adaptation	Dataset			Bookcorpus2 Val			Bookcorpus2 Train			Github val			Enron Val			Average		
		$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$
RMIA (shadow)	Prefix Tuning	1.00	0.89	0.56	1.00	0.90	0.55	1.00	0.93	0.63	1.00	0.88	0.58	1.00	0.90	0.58	1.00	0.90	0.58
	LoRA	1.00	0.70	0.52	1.00	0.69	0.53	1.00	0.74	0.52	1.00	0.73	0.52	1.00	0.71	0.52	1.00	0.71	0.52
	Full Fine-Tune	1.00	0.75	0.77	1.00	0.75	0.76	1.00	0.78	0.80	1.00	0.91	0.66	1.00	0.80	0.75	1.00	0.80	0.75
	Head Fine-Tune	1.00	0.72	0.73	1.00	0.72	0.72	1.00	0.80	0.74	1.00	0.57	0.65	1.00	0.70	0.71	1.00	0.70	0.71
	Average	1.00	0.77	0.65	1.00	0.76	0.64	1.00	0.81	0.67	1.00	0.77	0.60	1.00	0.78	0.64	1.00	0.78	0.64
RMIA (Pythia 1B)	Prefix Tuning	0.91	0.56	0.51	0.97	0.57	0.50	0.96	0.54	0.52	0.98	0.54	0.51	0.95	0.55	0.51	0.95	0.55	0.51
	LoRA	0.87	0.52	0.52	0.96	0.51	0.51	0.91	0.51	0.50	0.98	0.56	0.51	0.93	0.52	0.51	0.93	0.52	0.51
	Full Fine-Tune	0.99	0.54	0.52	1.00	0.54	0.52	0.99	0.53	0.52	0.99	0.59	0.50	1.00	0.55	0.51	1.00	0.55	0.51
	Head Fine-Tune	0.96	0.57	0.52	0.99	0.56	0.51	0.99	0.65	0.52	1.00	0.54	0.50	0.99	0.58	0.51	0.99	0.58	0.51
	Average	0.94	0.55	0.52	0.98	0.55	0.51	0.96	0.56	0.51	0.99	0.56	0.51	0.97	0.55	0.51	0.97	0.55	0.51
Reference (Pythia 1B)	Prefix Tuning	0.93	0.56	0.52	0.97	0.57	0.50	0.97	0.53	0.51	0.97	0.54	0.50	0.96	0.55	0.51	0.96	0.55	0.51
	LoRA	0.89	0.52	0.52	0.97	0.51	0.51	0.92	0.51	0.50	0.97	0.55	0.51	0.94	0.52	0.51	0.94	0.52	0.51
	Full Fine-Tune	1.00	0.54	0.52	1.00	0.54	0.52	0.99	0.54	0.52	0.98	0.59	0.50	0.99	0.55	0.51	0.99	0.55	0.51
	Head Fine-Tune	0.98	0.57	0.52	1.00	0.56	0.51	0.99	0.66	0.50	0.99	0.54	0.50	0.99	0.58	0.51	0.99	0.58	0.51
	Average	0.95	0.55	0.52	0.98	0.55	0.51	0.97	0.56	0.51	0.98	0.55	0.50	0.97	0.55	0.51	0.97	0.55	0.51
Min-K%	Prefix Tuning	0.78	0.51	0.50	0.70	0.51	0.50	0.65	0.52	0.52	0.66	0.51	0.52	0.70	0.51	0.51	0.70	0.51	0.51
	LoRA	0.67	0.51	0.51	0.63	0.50	0.50	0.61	0.52	0.52	0.65	0.51	0.51	0.64	0.51	0.51	0.64	0.51	0.51
	Full Fine-Tune	0.87	0.51	0.51	0.82	0.50	0.50	0.77	0.52	0.52	0.78	0.51	0.51	0.81	0.51	0.51	0.81	0.51	0.51
	Head Fine-Tune	0.75	0.51	0.51	0.72	0.50	0.51	0.64	0.52	0.52	0.70	0.51	0.51	0.70	0.51	0.51	0.70	0.51	0.51
	Average	0.77	0.51	0.51	0.72	0.50	0.50	0.67	0.52	0.52	0.70	0.51	0.51	0.71	0.51	0.51	0.71	0.51	0.51

Table 5: **Membership Inference for OOD Adaptations using Pythia 1.4B.** We present the AUC scores obtained with reference, and Min-K% MIAs for the Pythia 1.4B model adapted on different datasets with  $\varepsilon \in \{0.1, 8, \infty\}$ .

MIA	Adaptation	Dataset			Samsun			German Wiki			Average		
		$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$
RMIA (shadow)	Prefix	0.58	0.77	0.54	1.00	0.85	0.56	0.79	0.81	0.55	0.79	0.81	0.55
	LoRA	0.53	0.79	0.51	1.00	0.82	0.64	0.76	0.81	0.58	0.76	0.81	0.58
	Full Fine-Tune	1.00	0.99	0.62	1.00	1.00	0.90	1.00	1.00	0.76	1.00	1.00	0.76
	Head Fine-Tune	0.95	1.00	0.85	1.00	0.90	0.89	0.97	0.95	0.87	0.97	0.95	0.87
	Average	0.76	0.94	0.63	1.00	0.89	0.75	0.88	0.89	0.69	0.88	0.89	0.69
RMIA (Pythia 1B)	Prefix	0.52	0.52	0.51	0.92	0.53	0.50	0.72	0.53	0.51	0.72	0.53	0.51
	LoRA	0.50	0.54	0.50	0.97	0.51	0.50	0.74	0.52	0.50	0.74	0.52	0.50
	Full Fine-Tune	1.00	0.52	0.50	1.00	0.58	0.51	1.00	0.55	0.51	1.00	0.55	0.51
	Head Fine-Tune	0.51	0.56	0.51	0.92	0.61	0.52	0.71	0.59	0.51	0.71	0.59	0.51
	Average	0.63	0.54	0.50	0.95	0.56	0.51	0.79	0.55	0.51	0.79	0.55	0.51
Reference (Pythia 1B)	Prefix	0.52	0.52	0.51	0.93	0.54	0.49	0.72	0.53	0.50	0.72	0.53	0.50
	LoRA	0.50	0.53	0.50	0.98	0.51	0.49	0.74	0.52	0.49	0.74	0.52	0.49
	Full Fine-Tune	1.00	0.52	0.50	1.00	0.59	0.51	1.00	0.55	0.51	1.00	0.55	0.51
	Head Fine-Tune	0.51	0.56	0.51	0.93	0.61	0.51	0.72	0.59	0.51	0.72	0.59	0.51
	Average	0.63	0.53	0.50	0.96	0.56	0.50	0.80	0.55	0.50	0.80	0.55	0.50
Min-K%	Prefix	0.52	0.51	0.51	0.70	0.53	0.50	0.61	0.52	0.51	0.61	0.52	0.51
	LoRA	0.50	0.52	0.50	0.79	0.52	0.51	0.65	0.52	0.51	0.65	0.52	0.51
	Full Fine-Tune	1.00	0.51	0.51	0.98	0.54	0.52	0.99	0.53	0.51	0.99	0.53	0.51
	Head Fine-Tune	0.51	0.53	0.51	0.74	0.55	0.52	0.62	0.54	0.52	0.62	0.54	0.52
	Average	0.63	0.52	0.51	0.80	0.53	0.51	0.72	0.53	0.51	0.72	0.53	0.51

Table 6: **Membership Inference for IID Adaptations using Pythia 1.4B.** We present the AUC scores obtained with reference, and Min-K% MIAs for the Pythia 1.4B model adapted on different datasets with  $\varepsilon \in \{0.1, 8, \infty\}$ .

MIA	Adaptation	Dataset			Pile Bookcorpus2 Val			Pile Bookcorpus2 Train			Pile Github Val			Pile Enron Val			Average		
		$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$
RMIA (shadow)	Prefix	1.00	0.68	0.54	1.00	0.68	0.55	1.00	0.62	0.56	1.00	0.72	0.60	1.00	0.68	0.56	1.00	0.68	0.56
	LoRA	0.96	0.99	0.51	0.74	0.98	0.52	1.00	0.97	0.53	1.00	0.99	0.67	0.93	0.98	0.56	0.93	0.98	0.56
	Full Fine-Tune	0.98	1.00	0.71	0.99	0.99	0.70	1.00	0.99	0.71	1.00	1.00	0.62	0.99	0.99	0.69	0.99	0.99	0.69
	Head Fine-Tune	1.00	1.00	0.72	1.00	1.00	0.69	1.00	1.00	0.71	1.00	1.00	0.64	1.00	1.00	0.69	1.00	1.00	0.69
	Average	0.99	0.92	0.62	0.93	0.92	0.62	1.00	0.89	0.63	1.00	0.93	0.63	0.98	0.91	0.62	0.98	0.91	0.62
RMIA (Pythia 1B)	Prefix	0.79	0.52	0.51	0.85	0.52	0.51	0.76	0.51	0.51	0.78	0.51	0.51	0.79	0.52	0.51	0.79	0.52	0.51
	LoRA	0.56	0.58	0.51	0.50	0.59	0.51	0.90	0.57	0.52	0.97	0.59	0.51	0.73	0.58	0.51	0.73	0.58	0.51
	Full Fine-Tune	0.64	0.59	0.51	0.65	0.58	0.50	0.97	0.55	0.50	0.99	0.57	0.51	0.81	0.57	0.51	0.81	0.57	0.51
	Head Fine-Tune	0.79	0.64	0.50	0.54	0.63	0.50	0.91	0.64	0.51	0.99	0.64	0.51	0.81	0.64	0.51	0.81	0.64	0.51
	Average	0.69	0.58	0.51	0.64	0.58	0.50	0.88	0.57	0.51	0.93	0.58	0.51	0.79	0.58	0.51	0.79	0.58	0.51
Reference (Pythia 1B)	Prefix	0.80	0.52	0.51	0.86	0.52	0.51	0.76	0.50	0.50	0.77	0.49	0.50	0.80	0.51	0.50	0.80	0.51	0.50
	LoRA	0.57	0.58	0.51	0.49	0.59	0.51	0.92	0.55	0.50	0.96	0.60	0.50	0.73	0.58	0.51	0.73	0.58	0.51
	Full Fine-Tune	0.64	0.58	0.51	0.65	0.57	0.49	0.98	0.53	0.50	0.99	0.58	0.51	0.81	0.56	0.50	0.81	0.56	0.50
	Head Fine-Tune	0.80	0.64	0.51	0.54	0.64	0.50	0.91	0.67	0.51	0.99	0.65	0.50	0.81	0.65	0.50	0.81	0.65	0.50
	Average	0.70	0.58	0.51	0.64	0.58	0.50	0.89	0.56	0.50	0.93	0.58	0.50	0.79	0.58	0.50	0.79	0.58	0.50
Min-K%	Prefix	0.61	0.50	0.49	0.58	0.50	0.50	0.57	0.52	0.51	0.57	0.51	0.51	0.58	0.51	0.51	0.58	0.51	0.51
	LoRA	0.50	0.51	0.50	0.50	0.51	0.50	0.64	0.53	0.52	0.68	0.52	0.51	0.58	0.52	0.51	0.58	0.52	0.51
	Full Fine-Tune	0.52	0.53	0.49	0.52	0.54	0.50	0.73	0.56	0.51	0.83	0.52	0.51	0.65	0.54	0.51	0.65	0.54	0.51
	Head Fine-Tune	0.57	0.53	0.49	0.51	0.53	0.50	0.61	0.54	0.51	0.90	0.53	0.51	0.64	0.53	0.51	0.64	0.53	0.51
	Average	0.55	0.52	0.50	0.53	0.52	0.50	0.64	0.54	0.52	0.75	0.52	0.51	0.61	0.52	0.51	0.61	0.52	0.51

## 741 C.2 Exposure

742 Table 17 and Table 18 show the exposure performance of the four types of canary prefixes. With  
743 canary exposure, we do not use any shadow or reference models. Therefore, the results are often close  
744 to random guessing when using DP for LLM adaptations. However, the results for canary exposure  
745 are still much higher than for Min-K%, the closest MIA method executed with the same assumptions.

Table 7: **Membership Inference for OOD Adaptations using Pythia 410M.** We present the AUC scores obtained with reference, and Min-K% MIAs for the Pythia 410M model adapted on different datasets with  $\varepsilon \in \{0.1, 8, \infty\}$ .

MIA	Adaptation	Dataset			Samsun			German Wiki			Average		
		$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$
RMIA (shadow)	Prefix	0.87	0.67	0.51	0.90	0.66	0.50	0.88	0.67	0.51	0.88	0.67	0.51
	LoRA	0.93	0.62	0.52	0.71	0.97	0.54	0.82	0.79	0.53	0.82	0.79	0.53
	Full Fine-Tune	0.99	0.98	0.52	1.00	1.00	0.53	1.00	0.99	0.52	1.00	0.99	0.52
	Head Fine-Tune	1.00	0.76	0.76	0.94	1.00	0.82	0.97	0.88	0.79	0.97	0.88	0.79
	Average	0.95	0.76	0.58	0.89	0.91	0.60	0.92	0.83	0.59	0.92	0.83	0.59
RMIA (Pythia 1B)	Prefix	0.54	0.52	0.51	0.58	0.51	0.50	0.56	0.52	0.51	0.56	0.52	0.51
	LoRA	0.52	0.50	0.52	0.51	0.56	0.50	0.51	0.53	0.51	0.51	0.53	0.51
	Full Fine-Tune	0.80	0.55	0.50	0.93	0.58	0.51	0.86	0.56	0.50	0.86	0.56	0.50
	Head Fine-Tune	0.80	0.50	0.50	0.51	0.62	0.51	0.66	0.56	0.51	0.66	0.56	0.51
	Average	0.66	0.52	0.51	0.63	0.57	0.50	0.65	0.54	0.51	0.65	0.54	0.51
Reference (Pythia 1B)	Prefix	0.54	0.52	0.51	0.57	0.50	0.48	0.55	0.51	0.49	0.55	0.51	0.49
	LoRA	0.52	0.49	0.51	0.50	0.55	0.48	0.51	0.52	0.49	0.51	0.52	0.49
	Full Fine-Tune	0.79	0.55	0.50	0.92	0.56	0.49	0.85	0.55	0.49	0.85	0.55	0.49
	Head Fine-Tune	0.79	0.49	0.50	0.51	0.62	0.49	0.65	0.56	0.49	0.65	0.56	0.49
	Average	0.66	0.51	0.51	0.63	0.56	0.48	0.64	0.54	0.49	0.64	0.54	0.49
Min-K%	Prefix	0.52	0.51	0.51	0.54	0.52	0.51	0.53	0.52	0.51	0.53	0.52	0.51
	LoRA	0.51	0.50	0.51	0.52	0.54	0.51	0.51	0.52	0.51	0.51	0.52	0.51
	Full Fine-Tune	0.69	0.53	0.50	0.79	0.54	0.52	0.74	0.53	0.51	0.74	0.53	0.51
	Head Fine-Tune	0.69	0.50	0.50	0.52	0.56	0.52	0.60	0.53	0.51	0.60	0.53	0.51
	Average	0.60	0.51	0.51	0.59	0.54	0.51	0.60	0.53	0.51	0.60	0.53	0.51

Table 8: **Membership Inference for IID Adaptations using Pythia 410M.** We present the AUC scores obtained with reference, and Min-K% MIAs for the Pythia 410M model adapted on different datasets with  $\varepsilon \in \{0.1, 8, \infty\}$ .

MIA	Adaptation	Dataset			Pile Bookcorpus2 Val			Pile Bookcorpus2 Train			Pile Github Val			Pile Enron Val			Average		
		$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$
RMIA (shadow)	Prefix	0.83	0.65	0.55	0.86	0.67	0.52	0.65	0.69	0.51	0.89	0.65	0.53	0.81	0.67	0.53	0.81	0.67	0.53
	LoRA	0.72	0.91	0.58	0.73	0.89	0.57	0.74	0.92	0.51	0.74	0.98	0.57	0.73	0.92	0.56	0.73	0.92	0.56
	Full Fine-Tune	1.00	1.00	0.60	1.00	1.00	0.57	1.00	0.98	0.51	0.99	0.98	0.66	1.00	0.99	0.58	1.00	0.99	0.58
	Head Fine-Tune	0.96	1.00	0.74	0.96	1.00	0.69	1.00	1.00	0.62	1.00	1.00	0.72	0.98	1.00	0.69	0.98	1.00	0.69
	Average	0.87	0.89	0.62	0.89	0.89	0.59	0.85	0.90	0.54	0.91	0.90	0.62	0.88	0.90	0.59	0.88	0.90	0.59
RMIA (Pythia 1B)	Prefix	0.56	0.51	0.51	0.56	0.52	0.52	0.53	0.52	0.51	0.53	0.51	0.51	0.54	0.52	0.51	0.54	0.52	0.51
	LoRA	0.50	0.55	0.51	0.50	0.55	0.51	0.51	0.54	0.50	0.50	0.54	0.51	0.50	0.54	0.51	0.50	0.54	0.51
	Full Fine-Tune	0.91	0.58	0.50	0.93	0.59	0.51	0.91	0.55	0.52	0.83	0.54	0.50	0.90	0.57	0.51	0.90	0.57	0.51
	Head Fine-Tune	0.51	0.62	0.50	0.51	0.62	0.52	0.90	0.59	0.52	0.91	0.58	0.49	0.71	0.60	0.51	0.71	0.60	0.51
	Average	0.62	0.57	0.50	0.63	0.57	0.51	0.71	0.55	0.51	0.69	0.54	0.50	0.66	0.56	0.51	0.66	0.56	0.51
Reference (Pythia 1B)	Prefix	0.56	0.51	0.51	0.55	0.52	0.51	0.51	0.50	0.49	0.52	0.50	0.49	0.54	0.51	0.50	0.54	0.51	0.50
	LoRA	0.51	0.55	0.51	0.51	0.54	0.51	0.50	0.52	0.49	0.47	0.52	0.50	0.50	0.53	0.50	0.50	0.53	0.50
	Full Fine-Tune	0.91	0.57	0.50	0.93	0.58	0.51	0.88	0.53	0.49	0.80	0.52	0.49	0.88	0.55	0.50	0.88	0.55	0.50
	Head Fine-Tune	0.51	0.62	0.51	0.51	0.62	0.52	0.87	0.59	0.50	0.88	0.58	0.49	0.70	0.60	0.51	0.70	0.60	0.51
	Average	0.62	0.56	0.51	0.63	0.57	0.51	0.69	0.53	0.49	0.67	0.53	0.49	0.65	0.55	0.50	0.65	0.55	0.50
Min-K%	Prefix	0.51	0.50	0.50	0.52	0.51	0.51	0.53	0.52	0.50	0.52	0.51	0.52	0.52	0.51	0.51	0.52	0.51	0.51
	LoRA	0.50	0.51	0.50	0.50	0.52	0.50	0.51	0.54	0.51	0.50	0.52	0.51	0.50	0.52	0.51	0.50	0.52	0.51
	Full Fine-Tune	0.82	0.52	0.50	0.80	0.53	0.50	0.74	0.56	0.52	0.75	0.54	0.49	0.78	0.54	0.50	0.78	0.54	0.50
	Head Fine-Tune	0.50	0.53	0.49	0.50	0.53	0.51	0.62	0.53	0.51	0.68	0.52	0.50	0.57	0.53	0.50	0.57	0.53	0.50
	Average	0.58	0.52	0.50	0.58	0.52	0.51	0.60	0.54	0.51	0.61	0.52	0.51	0.59	0.52	0.51	0.59	0.52	0.51

Table 9: **Membership Inference for OOD Adaptations using Pythia 160M.** We present the AUC scores obtained with reference, and Min-K% MIAs for the Pythia 160M model adapted on different datasets with  $\varepsilon \in \{0.1, 8, \infty\}$ .

MIA	Adaptation	Dataset			Samsun			German Wiki			Average		
		$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$
RMIA (shadow)	Prefix	0.55	0.53	0.52	0.56	0.62	0.53	0.56	0.57	0.53	0.56	0.57	0.53
	LoRA	0.78	0.61	0.55	0.62	0.57	0.59	0.70	0.59	0.57	0.70	0.59	0.57
	Full Fine-Tune	1.00	0.74	0.61	0.90	0.99	0.65	0.95	0.86	0.63	0.95	0.86	0.63
	Head Fine-Tune	1.00	0.89	0.73	0.96	0.75	0.77	0.98	0.82	0.75	0.98	0.82	0.75
	Average	0.83	0.69	0.60	0.76	0.73	0.64	0.80	0.71	0.62	0.80	0.71	0.62
RMIA (Pythia 1B)	Prefix	0.51	0.51	0.50	0.51	0.51	0.51	0.51	0.51	0.50	0.51	0.51	0.50
	LoRA	0.51	0.50	0.50	0.51	0.51	0.50	0.51	0.51	0.50	0.51	0.50	0.50
	Full Fine-Tune	0.81	0.52	0.50	0.52	0.55	0.50	0.66	0.53	0.50	0.66	0.53	0.50
	Head Fine-Tune	0.69	0.51	0.50	0.52	0.51	0.52	0.60	0.51	0.51	0.60	0.51	0.51
	Average	0.63	0.51	0.50	0.51	0.52	0.51	0.57	0.51	0.50	0.57	0.51	0.50
Reference (Pythia 1B)	Prefix	0.51	0.51	0.51	0.50	0.49	0.49	0.50	0.50	0.50	0.50	0.50	0.50
	LoRA	0.51	0.50	0.51	0.49	0.49	0.49	0.50	0.50	0.50	0.50	0.50	0.50
	Full Fine-Tune	0.79	0.52	0.50	0.50	0.53	0.49	0.65	0.52	0.49	0.65	0.52	0.49
	Head Fine-Tune	0.69	0.51	0.50	0.50	0.49	0.50	0.60	0.50	0.50	0.60	0.50	0.50
	Average	0.63	0.51	0.50	0.50	0.50	0.49	0.56	0.51	0.50	0.56	0.51	0.50
Min-K%	Prefix	0.51	0.51	0.51	0.52	0.52	0.51	0.51	0.51	0.51	0.51	0.51	0.51
	LoRA	0.51	0.50	0.50	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51
	Full Fine-Tune	0.71	0.52	0.50	0.52	0.53	0.51	0.61	0.52	0.51	0.61	0.52	0.51
	Head Fine-Tune	0.63	0.51	0.50	0.52	0.51	0.52	0.58	0.51	0.51	0.58	0.51	0.51
	Average	0.59	0.51	0.50	0.52	0.52	0.51	0.55	0.51	0.51	0.55	0.51	0.51



Table 10: **Membership Inference for IID Adaptations using Pythia 160M.** We present the AUC scores obtained with reference, and Min-K% MIAs for the Pythia 160M model adapted on different datasets with  $\varepsilon \in \{0.1, 8, \infty\}$ .

MIA	Adaptation	Dataset: Pile Bookcorpus2 Val			Pile Bookcorpus2 Train			Pile Github Val			Pile Enron Val			Average		
		$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$
RMIA (shadow)	Prefix	0.61	0.72	0.53	0.61	0.71	0.54	0.57	0.67	0.51	0.66	0.75	0.54	0.61	0.71	0.53
	LoRA	0.82	0.60	0.54	0.83	0.79	0.55	0.80	0.82	0.53	0.91	0.61	0.53	0.84	0.71	0.54
	Full Fine-Tune	1.00	0.89	0.58	0.89	0.93	0.56	1.00	0.95	0.56	1.00	0.97	0.52	0.97	0.94	0.55
	Head Fine-Tune	1.00	0.74	0.75	1.00	0.97	0.72	1.00	0.99	0.62	1.00	0.80	0.70	1.00	0.87	0.70
	Average	0.86	0.74	0.60	0.83	0.85	0.59	0.84	0.86	0.55	0.89	0.78	0.57	0.86	0.81	0.58
RMIA (Pythia 1B)	Prefix	0.50	0.50	0.50	0.52	0.53	0.52	0.51	0.51	0.51	0.50	0.50	0.50	0.51	0.51	0.51
	LoRA	0.50	0.50	0.50	0.51	0.52	0.51	0.51	0.51	0.51	0.50	0.50	0.50	0.51	0.51	0.51
	Full Fine-Tune	1.00	0.53	0.50	0.52	0.54	0.50	0.85	0.52	0.51	0.99	0.53	0.50	0.84	0.53	0.51
	Head Fine-Tune	0.71	0.50	0.50	0.74	0.56	0.51	0.67	0.55	0.51	0.78	0.50	0.50	0.73	0.53	0.51
	Average	0.68	0.51	0.50	0.57	0.53	0.51	0.64	0.52	0.51	0.69	0.51	0.50	0.65	0.52	0.51
Reference (Pythia 1B)	Prefix	0.50	0.51	0.50	0.52	0.52	0.52	0.50	0.50	0.50	0.50	0.50	0.50	0.51	0.51	0.50
	LoRA	0.51	0.50	0.50	0.52	0.52	0.51	0.50	0.51	0.50	0.49	0.49	0.49	0.50	0.50	0.50
	Full Fine-Tune	1.00	0.53	0.50	0.52	0.53	0.51	0.81	0.52	0.50	0.96	0.51	0.50	0.82	0.52	0.50
	Head Fine-Tune	0.69	0.50	0.50	0.71	0.55	0.51	0.64	0.54	0.50	0.72	0.49	0.49	0.69	0.52	0.50
	Average	0.68	0.51	0.50	0.57	0.53	0.51	0.61	0.52	0.50	0.67	0.50	0.49	0.63	0.51	0.50
Min-K%	Prefix	0.50	0.50	0.50	0.51	0.51	0.51	0.52	0.52	0.51	0.50	0.50	0.50	0.51	0.51	0.51
	LoRA	0.50	0.50	0.50	0.51	0.51	0.50	0.52	0.52	0.51	0.50	0.50	0.50	0.51	0.50	0.50
	Full Fine-Tune	0.96	0.51	0.50	0.51	0.51	0.50	0.67	0.52	0.52	0.93	0.52	0.51	0.77	0.52	0.50
	Head Fine-Tune	0.62	0.50	0.50	0.62	0.52	0.51	0.60	0.53	0.51	0.72	0.50	0.50	0.64	0.51	0.50
	Average	0.64	0.50	0.50	0.54	0.51	0.51	0.58	0.52	0.51	0.66	0.50	0.50	0.61	0.51	0.50

Table 11: **Membership Inference for OOD Adaptations using Pythia 70M.** We present the AUC scores obtained with reference, and Min-K% MIAs for the Pythia 70M model adapted on different datasets with  $\varepsilon \in \{0.1, 8, \infty\}$ .

MIA	Adaptation	Dataset: Samsun			German Wiki			Average		
		$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$
RMIA (shadow)	Prefix	0.53	0.62	0.51	0.60	0.63	0.56	0.57	0.63	0.53
	LoRA	0.68	0.58	0.55	0.59	0.61	0.57	0.63	0.59	0.56
	Full Fine-Tune	0.98	0.92	0.63	0.98	0.97	0.71	0.98	0.94	0.67
	Head Fine-Tune	1.00	0.93	0.73	0.95	0.93	0.77	0.97	0.93	0.75
	Average	0.80	0.76	0.61	0.78	0.78	0.65	0.79	0.77	0.63
RMIA (Pythia 1B)	Prefix	0.51	0.51	0.51	0.50	0.51	0.50	0.51	0.51	0.51
	LoRA	0.51	0.51	0.52	0.51	0.51	0.51	0.51	0.51	0.51
	Full Fine-Tune	0.52	0.53	0.52	0.53	0.55	0.50	0.53	0.54	0.51
	Head Fine-Tune	0.67	0.54	0.50	0.52	0.54	0.51	0.59	0.54	0.51
	Average	0.55	0.52	0.51	0.51	0.53	0.51	0.53	0.53	0.51
Reference (Pythia 1B)	Prefix	0.51	0.52	0.51	0.49	0.50	0.49	0.50	0.51	0.50
	LoRA	0.51	0.51	0.52	0.50	0.50	0.50	0.50	0.51	0.51
	Full Fine-Tune	0.52	0.53	0.52	0.51	0.53	0.49	0.52	0.53	0.51
	Head Fine-Tune	0.67	0.55	0.51	0.50	0.52	0.50	0.59	0.53	0.50
	Average	0.55	0.53	0.51	0.50	0.51	0.49	0.53	0.52	0.50
Min-K%	Prefix	0.51	0.51	0.50	0.51	0.52	0.51	0.51	0.51	0.51
	LoRA	0.51	0.50	0.51	0.52	0.52	0.52	0.51	0.51	0.51
	Full Fine-Tune	0.51	0.52	0.51	0.53	0.54	0.52	0.52	0.53	0.51
	Head Fine-Tune	0.64	0.53	0.50	0.53	0.54	0.52	0.58	0.54	0.51
	Average	0.54	0.52	0.50	0.52	0.53	0.52	0.53	0.52	0.51

Table 12: **Membership Inference for IID Adaptations using Pythia 70M.** We present the AUC scores obtained with reference, and Min-K% MIAs for the Pythia 70M model adapted on different datasets with  $\varepsilon \in \{0.1, 8, \infty\}$ .

MIA	Adaptation	Dataset: Pile Bookcorpus2 Val			Pile Bookcorpus2 Train			Pile Github Val			Pile Enron Val			Average		
		$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$
RMIA (shadow)	Prefix	0.63	0.66	0.50	0.60	0.68	0.52	0.57	0.73	0.56	0.65	0.72	0.53	0.61	0.70	0.53
	LoRA	0.57	0.80	0.54	0.81	0.80	0.55	0.84	0.83	0.55	0.85	0.63	0.50	0.77	0.77	0.54
	Full Fine-Tune	0.99	0.92	0.59	0.99	0.92	0.59	1.00	0.97	0.58	0.99	0.97	0.58	0.99	0.95	0.58
	Head Fine-Tune	0.97	0.94	0.72	1.00	0.95	0.70	1.00	0.98	0.76	1.00	0.97	0.76	0.99	0.96	0.73
	Average	0.79	0.83	0.59	0.85	0.84	0.59	0.85	0.88	0.61	0.87	0.82	0.59	0.84	0.84	0.60
RMIA (Pythia 1B)	Prefix	0.50	0.50	0.50	0.51	0.52	0.51	0.51	0.51	0.51	0.50	0.50	0.50	0.51	0.51	0.51
	LoRA	0.50	0.50	0.50	0.51	0.51	0.51	0.51	0.51	0.51	0.50	0.50	0.49	0.51	0.51	0.50
	Full Fine-Tune	0.52	0.52	0.50	0.53	0.53	0.51	0.87	0.52	0.51	0.51	0.52	0.50	0.61	0.52	0.51
	Head Fine-Tune	0.51	0.52	0.50	0.69	0.54	0.50	0.64	0.55	0.51	0.73	0.52	0.50	0.64	0.53	0.50
	Average	0.51	0.51	0.50	0.56	0.52	0.51	0.63	0.52	0.51	0.56	0.51	0.50	0.56	0.52	0.50
Reference (Pythia 1B)	Prefix	0.50	0.50	0.50	0.51	0.51	0.51	0.49	0.50	0.49	0.50	0.50	0.50	0.50	0.50	0.50
	LoRA	0.50	0.51	0.50	0.51	0.51	0.51	0.50	0.50	0.50	0.49	0.49	0.49	0.50	0.50	0.50
	Full Fine-Tune	0.52	0.52	0.50	0.52	0.53	0.51	0.81	0.51	0.50	0.50	0.51	0.49	0.59	0.52	0.50
	Head Fine-Tune	0.51	0.52	0.50	0.66	0.54	0.51	0.60	0.54	0.50	0.67	0.51	0.48	0.61	0.53	0.50
	Average	0.51	0.51	0.50	0.55	0.52	0.51	0.60	0.51	0.50	0.54	0.50	0.49	0.55	0.51	0.50
Min-K%	Prefix	0.50	0.50	0.49	0.51	0.51	0.51	0.52	0.52	0.52	0.52	0.52	0.52	0.51	0.51	0.51
	LoRA	0.49	0.50	0.50	0.51	0.51	0.51	0.52	0.52	0.52	0.51	0.51	0.51	0.51	0.51	0.51
	Full Fine-Tune	0.50	0.51	0.49	0.52	0.52	0.51	0.75	0.53	0.52	0.52	0.52	0.51	0.57	0.52	0.51
	Head Fine-Tune	0.50	0.51	0.49	0.64	0.53	0.51	0.62	0.55	0.51	0.76	0.53	0.50	0.63	0.53	0.50
	Average	0.50	0.50	0.49	0.54	0.52	0.51	0.60	0.53	0.52	0.58	0.52	0.51	0.55	0.52	0.51

### 746 C.3 Influence of Subset Size and Complexity

747 We evaluate how subset characteristics, specifically size and complexity (as measured by the perplexity  
748 in Table 2 in the original publication on the Pile [18]), affect privacy leakage. Specifically, for this



Table 13: **Membership Inference for OOD Adaptations using GPT Neo 1.3B.** We present the AUC scores obtained with reference, and Min-K% MIAs for the GPT Neo 1.3B model adapted on different datasets with  $\varepsilon \in \{0.1, 8, \infty\}$ .

MIA	Adaptation	Dataset			Samsun			German Wiki			Average		
		$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$
RMIA (shadow)	Prefix	0.50	0.51	0.50	0.98	0.50	0.51	0.74	0.50	0.51	0.74	0.50	0.51
	LoRA	0.53	0.85	0.51	0.55	0.89	0.50	0.54	0.87	0.51	0.54	0.87	0.51
	Full Fine-Tune	1.00	1.00	0.80	1.00	1.00	0.83	1.00	1.00	0.82	1.00	1.00	0.82
	Head Fine-Tune	0.93	1.00	0.81	0.93	1.00	0.85	0.93	1.00	0.83	0.93	1.00	0.83
	Average	0.74	0.84	0.66	0.86	0.85	0.68	0.80	0.84	0.67	0.80	0.84	0.67
RMIA (Pythia 1B)	Prefix	0.51	0.51	0.51	0.71	0.50	0.49	0.61	0.51	0.50	0.61	0.51	0.50
	LoRA	0.50	0.50	0.50	0.51	0.52	0.51	0.51	0.51	0.51	0.51	0.51	0.51
	Full Fine-Tune	0.58	0.56	0.51	0.74	0.63	0.51	0.66	0.60	0.51	0.66	0.60	0.51
	Head Fine-Tune	0.50	0.55	0.50	0.51	0.60	0.51	0.51	0.58	0.51	0.51	0.58	0.51
	Average	0.53	0.53	0.51	0.61	0.56	0.51	0.57	0.55	0.51	0.57	0.55	0.51
Reference (Pythia 1B)	Prefix	0.50	0.49	0.49	0.62	0.48	0.47	0.56	0.49	0.48	0.56	0.49	0.48
	LoRA	0.49	0.51	0.49	0.51	0.52	0.51	0.50	0.52	0.50	0.50	0.52	0.50
	Full Fine-Tune	0.59	0.57	0.49	0.74	0.61	0.49	0.66	0.59	0.49	0.66	0.59	0.49
	Head Fine-Tune	0.50	0.56	0.49	0.51	0.60	0.50	0.50	0.58	0.50	0.50	0.58	0.50
	Average	0.52	0.53	0.49	0.59	0.55	0.49	0.56	0.54	0.49	0.56	0.54	0.49
Min-K%	Prefix	0.52	0.50	0.50	0.65	0.50	0.51	0.58	0.50	0.51	0.58	0.50	0.51
	LoRA	0.51	0.51	0.51	0.52	0.53	0.52	0.52	0.52	0.52	0.52	0.52	0.52
	Full Fine-Tune	0.55	0.55	0.51	0.59	0.57	0.53	0.57	0.56	0.52	0.57	0.56	0.52
	Head Fine-Tune	0.51	0.54	0.51	0.53	0.56	0.52	0.52	0.55	0.52	0.52	0.55	0.52
	Average	0.52	0.52	0.51	0.57	0.54	0.52	0.55	0.53	0.51	0.55	0.53	0.51

Table 14: **Membership Inference for IID Adaptations using GPT Neo 1.3B.** We present the AUC scores obtained with reference, and Min-K% MIAs for the GPT Neo 1.3B model adapted on different datasets with  $\varepsilon \in \{0.1, 8, \infty\}$ .

MIA	Adaptation	Dataset			Pile Bookcorpus2 Val			Pile Bookcorpus2 Train			Pile Github Val			Pile Enron Val			Average		
		$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$
RMIA (shadow)	Prefix	0.62	0.51	0.52	0.78	0.51	0.50	0.95	0.50	0.51	0.79	0.57	0.55	0.78	0.52	0.52	0.78	0.52	0.52
	LoRA	0.53	0.81	0.52	0.54	0.81	0.51	0.55	0.89	0.50	0.57	0.64	0.62	0.55	0.79	0.54	0.55	0.79	0.54
	Full Fine-Tune	1.00	1.00	0.65	1.00	1.00	0.64	1.00	0.71	0.75	1.00	1.00	0.62	1.00	1.00	0.62	1.00	0.93	0.67
	Head Fine-Tune	0.96	1.00	0.70	1.00	1.00	0.70	1.00	1.00	0.87	1.00	1.00	0.65	0.99	1.00	0.73	0.99	1.00	0.73
	Average	0.78	0.83	0.60	0.83	0.83	0.59	0.87	0.78	0.66	0.84	0.80	0.61	0.83	0.81	0.61	0.83	0.81	0.61
RMIA (Pythia 1B)	Prefix	0.51	0.51	0.51	0.76	0.51	0.50	0.68	0.50	0.51	0.80	0.57	0.56	0.69	0.52	0.52	0.69	0.52	0.52
	LoRA	0.51	0.53	0.51	0.50	0.50	0.50	0.51	0.54	0.53	0.56	0.56	0.56	0.52	0.53	0.53	0.56	0.56	0.53
	Full Fine-Tune	0.72	0.62	0.51	0.71	0.62	0.51	0.91	0.54	0.53	0.70	0.67	0.57	0.76	0.61	0.53	0.76	0.61	0.53
	Head Fine-Tune	0.52	0.60	0.50	1.00	0.61	0.51	0.98	0.61	0.53	0.98	0.65	0.57	0.87	0.62	0.53	0.87	0.62	0.53
	Average	0.56	0.56	0.51	0.74	0.56	0.51	0.77	0.55	0.52	0.76	0.61	0.57	0.71	0.57	0.53	0.71	0.57	0.53
Reference (Pythia 1B)	Prefix	0.51	0.51	0.52	0.72	0.51	0.50	0.61	0.48	0.48	0.74	0.44	0.43	0.65	0.48	0.48	0.65	0.48	0.48
	LoRA	0.51	0.53	0.51	0.48	0.49	0.48	0.51	0.51	0.48	0.58	0.59	0.58	0.52	0.53	0.51	0.52	0.53	0.51
	Full Fine-Tune	0.72	0.62	0.52	0.71	0.62	0.51	0.89	0.50	0.50	0.74	0.65	0.56	0.77	0.60	0.52	0.77	0.60	0.52
	Head Fine-Tune	0.52	0.61	0.51	1.00	0.62	0.51	0.97	0.61	0.51	0.98	0.66	0.57	0.87	0.63	0.53	0.87	0.63	0.53
	Average	0.57	0.57	0.51	0.73	0.56	0.50	0.74	0.53	0.49	0.76	0.58	0.53	0.70	0.56	0.51	0.70	0.56	0.51
Min-K%	Prefix	0.50	0.51	0.51	0.65	0.52	0.50	0.67	0.52	0.53	0.77	0.55	0.55	0.65	0.53	0.52	0.65	0.53	0.52
	LoRA	0.50	0.50	0.50	0.50	0.50	0.50	0.52	0.54	0.53	0.57	0.57	0.57	0.52	0.53	0.52	0.52	0.53	0.52
	Full Fine-Tune	0.54	0.54	0.50	0.54	0.54	0.50	0.62	0.54	0.53	0.60	0.59	0.56	0.57	0.55	0.52	0.57	0.55	0.52
	Head Fine-Tune	0.50	0.52	0.49	0.91	0.53	0.50	0.74	0.54	0.53	0.87	0.59	0.57	0.76	0.55	0.52	0.76	0.55	0.52
	Average	0.51	0.52	0.50	0.65	0.52	0.50	0.64	0.54	0.53	0.70	0.58	0.56	0.63	0.54	0.52	0.63	0.54	0.52

Table 15: **Membership Inference for OOD Adaptations using GPT Neo 125M.** We present the AUC scores obtained with reference, and Min-K% MIAs for the GPT Neo 125M model adapted on different datasets with  $\varepsilon \in \{0.1, 8, \infty\}$ .

MIA	Adaptation	Dataset			Samsun			German Wiki			Average		
		$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$
RMIA (shadow)	Prefix	0.68	0.51	0.50	0.73	0.51	0.52	0.70	0.51	0.51	0.70	0.51	0.51
	LoRA	0.84	0.63	0.59	0.51	0.50	0.50	0.67	0.56	0.55	0.67	0.56	0.55
	Full Fine-Tune	1.00	0.99	0.72	1.00	0.52	0.79	1.00	0.75	0.75	1.00	0.75	0.75
	Head Fine-Tune	1.00	0.94	0.79	1.00	1.00	0.87	1.00	0.97	0.83	1.00	0.97	0.83
	Average	0.88	0.77	0.65	0.81	0.63	0.67	0.84	0.70	0.66	0.84	0.70	0.66
RMIA (Pythia 1B)	Prefix	0.52	0.51	0.51	0.55	0.50	0.50	0.54	0.50	0.50	0.54	0.50	0.50
	LoRA	0.51	0.51	0.51	0.50	0.50	0.50	0.50	0.51	0.51	0.50	0.51	0.51
	Full Fine-Tune	1.00	0.53	0.52	1.00	0.50	0.50	1.00	0.52	0.51	1.00	0.52	0.51
	Head Fine-Tune	1.00	0.51	0.50	0.54	0.57	0.51	0.77	0.54	0.51	0.77	0.54	0.51
	Average	0.76	0.52	0.51	0.65	0.52	0.50	0.70	0.52	0.51	0.70	0.52	0.51
Reference (Pythia 1B)	Prefix	0.52	0.49	0.49	0.51	0.48	0.48	0.51	0.49	0.49	0.51	0.49	0.49
	LoRA	0.51	0.51	0.51	0.49	0.49	0.49	0.50	0.50	0.50	0.50	0.50	0.50
	Full Fine-Tune	1.00	0.53	0.50	1.00	0.49	0.49	1.00	0.51	0.50	1.00	0.51	0.50
	Head Fine-Tune	1.00	0.51	0.50	0.53	0.55	0.49	0.76	0.53	0.50	0.76	0.53	0.50
	Average	0.76	0.51	0.50	0.63	0.50	0.49	0.69	0.51	0.49	0.69	0.51	0.49
Min-K%	Prefix	0.54	0.51	0.51	0.55	0.50	0.49	0.54	0.50	0.50	0.54	0.50	0.50
	LoRA	0.51	0.51	0.51	0.52	0.52	0.52	0.52	0.52	0.52	0.52	0.52	0.52
	Full Fine-Tune	1.00	0.53	0.52	1.00	0.52	0.52	1.00	0.53	0.52	1.00	0.53	0.52
	Head Fine-Tune	1.00	0.51	0.51	0.54	0.55	0.52	0.77	0.53	0.52	0.77	0.53	0.52
	Average	0.76	0.52	0.51	0.65	0.52	0.51	0.71	0.52	0.51	0.71	0.52	0.51

749 experiment, we use train subsets and adapt Pythia 1B privately with  $\varepsilon = 8$ . As shown in Figure 7, the

Table 16: **Membership Inference for IID Adaptations using GPT Neo 125M.** We present the AUC scores obtained with reference, and Min-K% MIAs for the GPT Neo 125M model adapted on different datasets with  $\varepsilon \in \{0.1, 8, \infty\}$ .

MIA	Adaptation	Dataset: Pile Bookcorpus2 Val			Pile Bookcorpus2 Train			Pile Github Val			Pile Enron Val			Average		
		$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$
RMIA (shadow)	Prefix	0.68	0.52	0.51	0.52	0.53	0.51	0.77	0.50	0.50	0.76	0.57	0.53	0.68	0.53	0.51
	LoRA	0.52	0.51	0.50	1.00	0.51	0.51	1.00	0.50	0.51	1.00	0.50	0.50	0.88	0.51	0.50
	Full Fine-Tune	1.00	0.51	0.68	1.00	0.97	0.58	0.98	0.97	0.68	1.00	0.98	0.66	1.00	0.86	0.65
	Head Fine-Tune	1.00	1.00	0.70	1.00	1.00	0.66	0.96	1.00	0.87	1.00	1.00	0.70	0.99	1.00	0.74
	Average	0.80	0.64	0.60	0.88	0.75	0.57	0.93	0.74	0.64	0.94	0.76	0.60	0.89	0.72	0.60
RMIA (Pythia 1B)	Prefix	0.52	0.50	0.50	0.52	0.51	0.50	0.54	0.53	0.53	0.55	0.54	0.54	0.54	0.52	0.52
	LoRA	0.50	0.50	0.50	0.94	0.51	0.51	0.72	0.52	0.52	0.90	0.56	0.56	0.77	0.52	0.52
	Full Fine-Tune	1.00	0.50	0.50	1.00	0.54	0.52	0.67	0.54	0.53	1.00	0.58	0.56	0.92	0.54	0.53
	Head Fine-Tune	1.00	0.56	0.50	1.00	0.57	0.51	0.95	0.57	0.53	1.00	0.59	0.56	0.99	0.57	0.52
	Average	0.75	0.52	0.50	0.87	0.53	0.51	0.72	0.54	0.53	0.86	0.57	0.56	0.80	0.54	0.52
Reference (Pythia 1B)	Prefix	0.52	0.51	0.51	0.52	0.51	0.50	0.50	0.48	0.48	0.55	0.43	0.43	0.52	0.48	0.48
	LoRA	0.51	0.51	0.51	0.92	0.51	0.51	0.70	0.50	0.50	0.87	0.53	0.53	0.75	0.51	0.51
	Full Fine-Tune	1.00	0.51	0.51	1.00	0.54	0.52	0.58	0.52	0.49	1.00	0.56	0.55	0.89	0.53	0.52
	Head Fine-Tune	1.00	0.56	0.51	1.00	0.57	0.51	0.92	0.55	0.51	1.00	0.57	0.53	0.98	0.56	0.51
	Average	0.76	0.52	0.51	0.86	0.53	0.51	0.67	0.51	0.50	0.85	0.52	0.51	0.79	0.52	0.51
Min-K%	Prefix	0.53	0.50	0.50	0.52	0.51	0.50	0.55	0.54	0.53	0.56	0.54	0.54	0.54	0.52	0.52
	LoRA	0.50	0.50	0.50	0.70	0.50	0.50	0.60	0.52	0.52	0.74	0.56	0.56	0.63	0.52	0.52
	Full Fine-Tune	1.00	0.50	0.50	1.00	0.53	0.51	0.80	0.55	0.53	1.00	0.58	0.56	0.95	0.54	0.53
	Head Fine-Tune	1.00	0.52	0.50	1.00	0.53	0.50	0.94	0.55	0.53	1.00	0.58	0.56	0.98	0.55	0.52
	Average	0.75	0.50	0.50	0.80	0.52	0.51	0.72	0.54	0.53	0.83	0.56	0.55	0.78	0.53	0.52

Table 17: **Canary Exposure for OOD Datasets.** Prefix Tuning and Full Fine-Tuning adaptation methods have a higher exposure on OOD datasets than the other adaptation approaches like LoRA and Head Fine-Tuning. We audit only the adaptations and assume the same pretrained LLM is used for all adaptations. We present the exposure scores obtained using the model loss for the Pythia 1B model adapted to different OOD datasets with  $\varepsilon \in \{0.1, 8, \infty\}$ . The exposure differs between the adaptations only for  $\varepsilon = \infty$  and approaches random guessing (values close to 1.44) for  $\varepsilon \in \{0.1, 8\}$ .

Canary Prefix Type	Adaptation	Dataset: SAMSum			German Wiki			Average		
		$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$
Random	Prefix Tuning	7.35	1.72	1.82	6.07	1.81	1.40	6.71	1.76	1.61
	LoRA	1.85	1.76	1.76	3.34	1.43	1.41	2.59	1.60	1.58
	Full Fine-Tune	6.91	1.77	1.75	5.76	1.43	1.43	6.33	1.60	1.59
	Head Fine-Tune	1.88	1.75	1.77	4.44	1.43	1.42	3.16	1.59	1.59
	Average	4.50	1.75	1.77	4.90	1.53	1.42	4.70	1.64	1.59
Rare	Prefix Tuning	6.44	1.41	1.55	5.22	1.82	2.11	5.83	1.61	1.83
	LoRA	1.54	1.49	1.52	2.47	1.81	1.79	2.01	1.65	1.66
	Full Fine-Tune	4.28	1.51	1.53	4.13	1.81	1.81	4.21	1.66	1.67
	Head Fine-Tune	1.54	1.56	1.52	3.65	1.81	1.80	2.60	1.69	1.66
	Average	3.45	1.49	1.53	3.87	1.81	1.88	3.66	1.65	1.70
Common	Prefix Tuning	7.54	1.97	1.81	5.02	2.17	2.54	6.28	2.07	2.17
	LoRA	1.90	1.92	2.00	2.84	1.75	1.82	2.37	1.83	1.91
	Full Fine-Tune	6.34	1.93	1.99	4.63	1.74	1.75	5.49	1.84	1.87
	Head Fine-Tune	3.05	1.93	1.98	3.30	1.74	1.76	3.18	1.83	1.87
	Average	4.71	1.94	1.94	3.95	1.85	1.97	4.33	1.89	1.96
Invisible	Prefix Tuning	5.16	2.14	2.19	7.17	1.96	1.25	6.16	2.05	1.72
	LoRA	3.82	1.74	1.61	2.54	1.44	1.40	3.18	1.59	1.50
	Full Fine-Tune	8.00	1.91	1.74	5.62	1.44	1.45	6.81	1.67	1.59
	Head Fine-Tune	5.91	1.67	1.59	3.66	1.44	1.45	4.78	1.55	1.52
	Average	5.72	1.87	1.78	4.75	1.57	1.39	5.23	1.72	1.58

Table 18: **Canary Exposure for IID Datasets.** We use the same setup as in Table 17 and observe the same trends, with higher privacy leakage for Prefix tuning and Full Fine-Tuning than for LoRA and Head Fine-Tuning.

Canary Prefix Type	Adaptation	Dataset: Bookcorpus2 Val			Bookcorpus2 Train			Github Val			Enron Val			Average		
		$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$
Random	Prefix Tuning	8.00	2.02	1.24	8.00	1.69	1.59	7.86	1.88	1.22	5.80	0.91	1.58	7.41	1.63	1.41
	LoRA	3.65	2.06	2.05	3.19	1.55	1.55	3.22	1.89	1.88	2.04	0.67	0.67	3.03	1.54	1.54
	Full Fine-Tune	6.59	2.04	4.00	6.45	1.60	3.88	6.52	1.91	3.07	4.38	0.70	4.00	5.98	1.56	3.74
	Head Fine-Tune	2.81	2.03	1.84	2.54	1.58	1.59	2.70	1.89	1.85	1.20	0.69	0.75	2.26	1.55	1.51
	Average	5.26	2.04	2.28	5.00	1.61	2.15	5.08	1.89	2.01	3.35	0.74	1.75	4.67	1.57	2.05
Rare	Prefix Tuning	8.00	1.39	0.93	7.94	1.39	2.06	7.79	1.60	1.17	6.13	1.15	1.93	7.47	1.38	1.52
	LoRA	3.24	1.54	1.54	2.48	1.30	1.30	2.31	1.67	1.67	2.15	1.24	1.23	2.55	1.44	1.44
	Full Fine-Tune	5.40	1.54	3.23	4.87	1.31	2.82	4.73	1.68	4.52	4.05	1.27	1.79	4.76	1.45	3.09
	Head Fine-Tune	2.64	1.53	1.46	1.97	1.30	1.45	2.18	1.67	1.54	1.73	1.22	1.10	2.13	1.43	1.39
	Average	4.32	1.50	1.79	4.32	1.32	1.91	4.25	1.65	2.23	3.52	1.22	1.51	4.23	1.42	1.86
Common	Prefix Tuning	6.61	1.44	2.29	7.05	1.71	2.09	6.79	1.60	2.50	5.08	0.86	2.36	6.38	1.40	2.31
	LoRA	3.83	1.58	1.59	3.56	1.72	1.72	3.81	1.75	1.75	2.15	0.89	0.89	3.33	1.49	1.49
	Full Fine-Tune	5.27	1.60	2.91	4.66	1.75	2.80	6.24	1.74	3.08	3.60	0.90	1.98	4.94	1.50	2.69
	Head Fine-Tune	1.68	1.57	1.40	1.85	1.74	1.60	2.28	1.74	1.64	1.15	0.92	0.87	1.74	1.49	1.37
	Average	4.35	1.55	2.04	4.28	1.73	2.05	4.78	1.71	2.24	2.90	0.89	1.52	4.10	1.47	1.97
Invisible	Prefix Tuning	2.45	1.10	1.54	2.22	1.45	1.63	6.41	1.47	1.55	0.88	1.76	2.07	2.90	1.45	1.70
	LoRA	3.93	1.30	1.30	4.02	1.41	1.40	3.68	1.27	1.26	0.77	0.80	0.80	3.10	1.19	1.19
	Full Fine-Tune	8.00	1.34	1.32	8.00	1.45	1.52	6.30	1.30	1.33	5.21	0.78	0.82	6.88	1.22	1.25
	Head Fine-Tune	1.96	1.29	1.29	2.01	1.40	1.41	2.01	1.24	1.27	1.48	0.80	0.80	1.87	1.18	1.19
	Average	4.08	1.26	1.36	4.06	1.43	1.49	4.60	1.32	1.35	2.69	1.03	1.12	3.71	1.26	1.33

analysis suggests that privacy leakage in datasets is influenced both by dataset size and the inherent complexity or diversity within the data. For instance, the largest subset with the CC dataset incurs the highest privacy leakage, likely due to its significant volume and potentially diverse content (with a perplexity of around 0.7). The other large and complex subsets, like ArXiv (a perplexity of around 0.77), also have high leakage levels. For ArXiv compared to Freelaw (which is similar in size but less diverse with a perplexity of around 0.6), ArXiv’s diversity increases leakage, as more unique samples

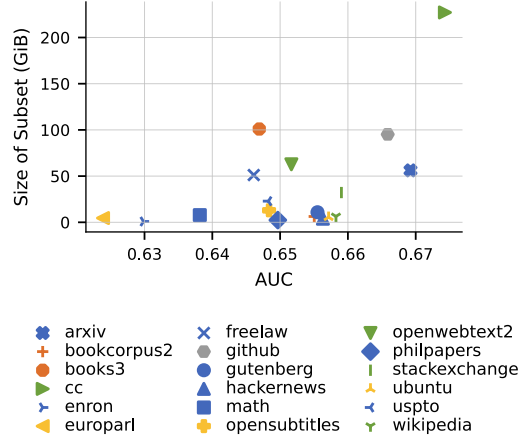


Figure 7: **Subset Size and Complexity.** The effect of the pretraining data subsets’ size and complexity on the incurred privacy leakage from the corresponding LLM adaptations. We evaluate the leakage using AUC and the Pythia 1B adapted with  $\varepsilon = 8$ .

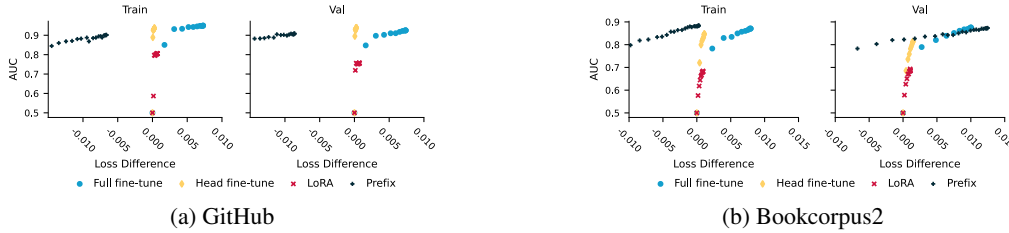


Figure 8: **Overlap (Train) and IID data (Val) show the same amount of privacy leakage across training.** The x-axis shows the difference between the initial pretrained loss and the evaluation loss. The y-axis represents the AUC score. We adapt Pythia 1B with  $\varepsilon = 8$ .

may need to be memorized. Finally, much smaller and more structured subsets like Europarl (with a perplexity of 0.75) and Enron Emails (the smallest subset) exhibit the least leakage, likely due to limited diversity and lower complexity.

#### C.4 Per-epoch Loss

We compare the development of AUC scores during training on IID and overlap data, as shown in Figure 8. These results display the AUC score at each epoch during training. To better compare IID and overlap data, we adjust the x-axis to represent the loss difference at each training step, calculated as the initial pretraining loss minus the adapted loss at the current training step. This calibration of the x-axis allows us to compare the two dataset types more precisely. With this setup, we evaluate two subsets of the Pile pretraining set: GitHub and BookCorpus2. First, the figures indicate that further adapting a model on IID data does not significantly improve its performance on that data, with the loss decreasing by only a maximum of 0.015 (GitHub with Full Fine-Tune). However, the observed increase in AUC score throughout training shows that the model does learn from the adaptation data.

#### C.5 Prefix Exposure

To investigate where privacy leakage comes from, we present the exposure observed with canary prefixes of varying lengths, after adapting Pythia 1B on the Github Val dataset with  $\varepsilon = \infty$ . Figure 9 show the exposure when only considering the first  $N$  tokens. This highlights that the prefix itself is the main source of privacy leakage.

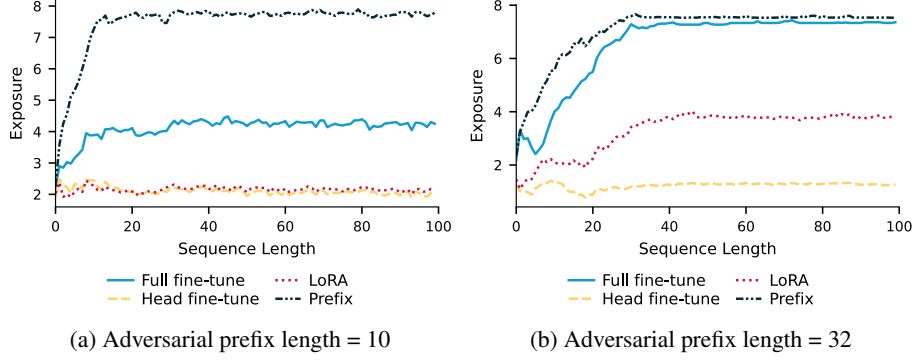


Figure 9: **The privacy leakage comes mostly from the adversarial prefix and much less from the interaction between the prefix and the sample.** We present the exposure when considering different lengths of canary prefixes after adapting Pythia 1B on Github Val. The evaluation was done for  $\varepsilon = \infty$ .

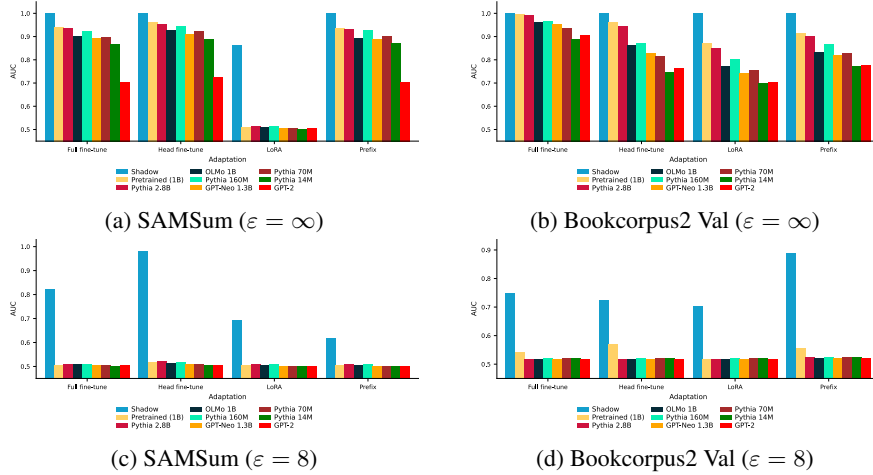


Figure 10: **Using at least one shadow model is crucial for RMIA, particularly for differentially private adaptations.** We present the AUC using RMIA with different types of shadow models after adapting Pythia 1B on Bookcorpus2 Val and SAMSum. The evaluation was done for  $\varepsilon = \{8, \infty\}$ .

## 774 D Influence of the Attacker’s Knowledge

775 We can observe how impactful an attacker’s knowledge about the target model and its pertaining  
 776 data is. Specifically, under moderate privacy regimes (*i.e.*,  $\varepsilon = 8$ ), *RMIA (shadow)* consistently  
 777 achieves best performance among models and datasets, as indicated in Table 5 - Table 16. However,  
 778 the effectiveness of MIAs quickly drops off when we move to more realistic scenarios, such as using  
 779 a pretrained model as a shadow model or having no shadow models available at all.

780 To model attackers with varying levels of background knowledge, we use a range of *shadow* models,  
 781 including Pythia 14M, Pythia 160M, Pythia 1B, Pythia 2.8B [3], GPT-neox [4], OLMo-1B [20],  
 782 and GPT-2 [40]. Therefore, we can simulate various attacker capabilities and assess their impact on  
 783 RMIA’s effectiveness. As we can see in Figure 10, the choice of reference model has a small impact  
 784 when attacking models fine-tuned on OOD data, even when architectural differences exist, such as  
 785 between GPT-Neo 1.3B and OLMo-1B. On the other hand, the MIA achieves higher success rates on  
 786 IID data when targeting the Pythia 1B model.

787 Additionally, Figure 11 illustrates the performance of various potential reference models over time.  
 788 We consistently observe the significant impact of knowing the target model’s architecture, especially  
 789 when the target and *shadow* models share the same architecture. The only exception to this pattern  
 790 appears in one of OOD datasets, SAMSum.

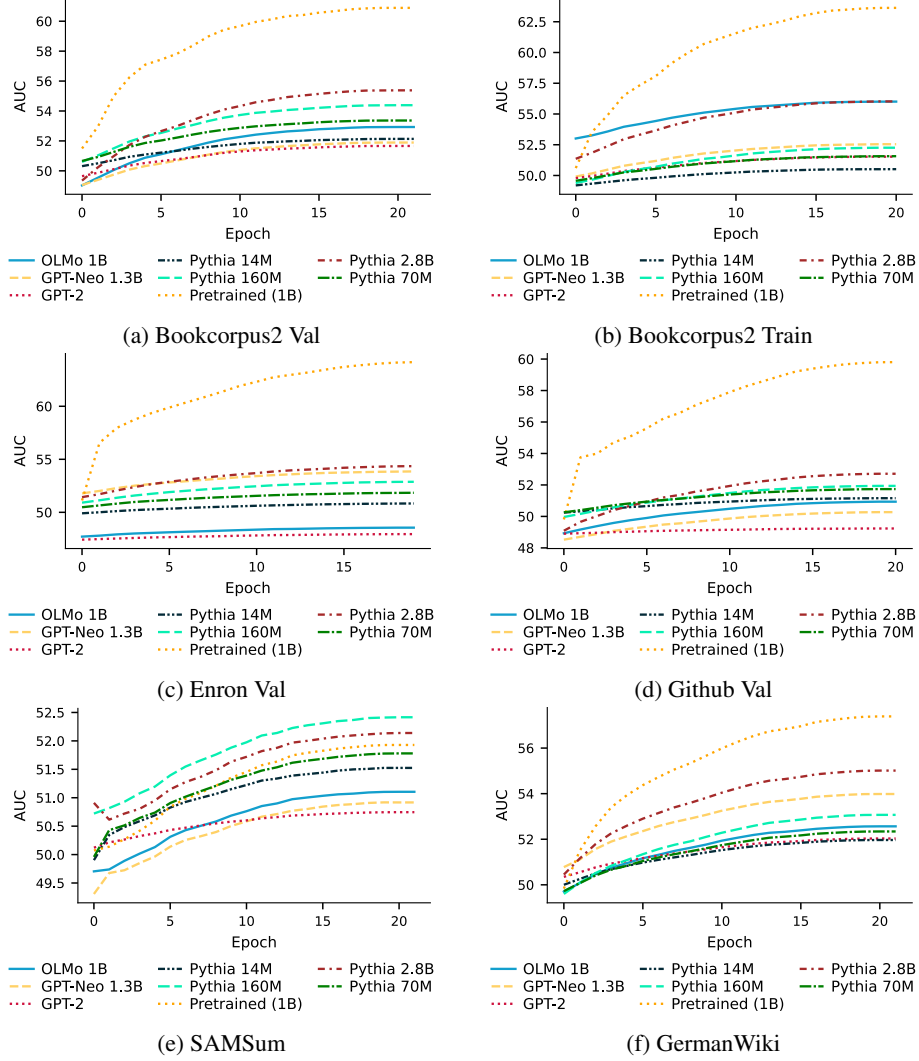


Figure 11: **Further analysis of the effectiveness of RMIA with pretrained models as a reference model.** As an extension of Figure 2, we fully fine-tuned Pythia 1B with  $\varepsilon = 8$  using three additional IID datasets: Bookcorpus2 Train, Github Train, and Github Val.

## E Loss Values

### E.1 Initial Loss of the LLM

Table 19 shows the loss at initialization for each dataset for the pretrained model and for a model adapted with an untrained Prefix Tuning.

Table 19: **Initial Losses for the Pythia 1B model on different datasets.** Standard refers to the model with default initialization, whereas Prefix refers to prepending an untrained Prefix Tuning to the hidden states.

Adaptation \ Dataset	SAMSum	GermanWiki	Bookcorpus2 Val	Bookcorpus2 Train	GitHub Val	Enron Val
	$\varepsilon = 0$	$\varepsilon = 0$	$\varepsilon = 0$	$\varepsilon = \infty$	$\varepsilon = 0$	$\varepsilon = 0$
Standard	2.747	2.732	3.011	2.997	1.539	2.388
Prefix Tuning	3.161	5.348	3.529	3.534	2.141	3.062

### E.2 Final Loss of the LLM

Table 20 show the final loss on the validation set. The hyperparameters are chosen to have similar loss between different adaptations using the same dataset and  $\varepsilon$ .

Table 20: **Validation loss values for the Pythia 1B model on different adaptation datasets.**

Adaptation \ Dataset	SAMSum			German Wiki			Bookcorpus2 Val			Bookcorpus2 Train			GitHub Val			Enron Val		
	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$	$\varepsilon = \infty$	$\varepsilon = 8$	$\varepsilon = 0.1$
Prefix Tuning	2.311	2.451	2.778	2.373	2.738	2.838	2.968	2.993	3.387	2.997	2.984	3.390	1.399	1.587	1.654	2.412	2.426	3.002
LoRA	2.313	2.462	2.761	2.378	2.737	2.801	2.991	3.007	3.013	2.979	3.002	3.003	1.558	1.572	1.558	2.394	2.402	2.403
Full Fine-Tune	2.251	2.457	2.739	2.511	2.726	2.747	2.934	2.999	3.028	2.960	2.995	3.020	1.598	1.577	1.575	2.375	2.397	2.413
Head Fine-Tune	2.354	2.454	2.761	2.574	2.731	2.756	2.949	3.007	3.139	2.966	3.002	3.132	1.577	1.573	1.570	2.409	2.403	2.436
Average	2.307	2.456	2.764	2.350	2.733	2.783	2.950	3.002	3.102	2.976	2.998	3.106	1.583	1.567	1.574	2.397	2.407	2.509

## F Exposure Estimation

There are two common ways to estimate the exposure [6]: (1) by sampling and (2) by distribution modeling. Figure 12 shows that the two approximations are similar when using 256 non-member samples. To statistically show the correlation, we use the Pearson correlation test, where the null hypothesis is that the distributions underlying the samples are uncorrelated and normally distributed. The data gives an extremely small p-value, which indicates a linear correlation between the two approximation methods.

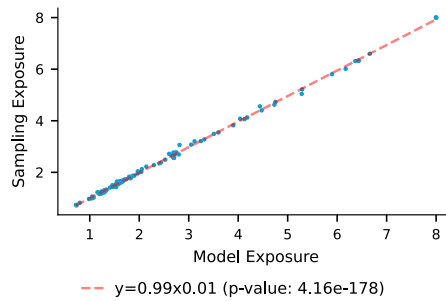


Figure 12: **The two ways to approximate the exposure are similar.** The relation between the model exposure and sampling exposure. The p-value is related to the Pearson correlation test.

## G Memorization of the Pretrained Model

Table 27 shows the number of memorized samples in the pretrained model.

Table 21: Validation loss values for the Pythia 1.4B model on different adaptation datasets.

Adaptation	Dataset	Samsum			German Wiki			Bookcorpus2 Val			Bookcorpus2 Train			Github Val			Enron Val		
		$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$
Prefix		2.712	2.456	3.451	2.465	2.655	5.246	2.901	3.538	3.857	2.929	3.657	3.918	1.542	2.564	2.909	2.411	2.973	3.791
LoRA		2.677	2.362	2.682	2.456	2.498	4.112	2.895	3.046	3.887	2.923	3.055	3.945	1.492	1.751	2.401	2.296	2.347	2.779
Full fine-tune		2.779	2.262	2.639	2.458	2.493	2.595	2.885	3.815	2.975	2.889	3.872	2.965	1.492	2.739	1.534	2.299	2.283	2.719
Head fine-tune		2.665	2.454	3.038	2.465	2.625	3.151	2.889	3.273	3.594	2.920	3.292	3.584	1.502	1.743	1.877	2.389	2.621	2.543
Average		2.708	2.384	2.952	2.461	2.568	3.776	2.892	3.418	3.578	2.915	3.469	3.603	1.507	2.199	2.180	2.349	2.556	2.858

Table 22: Validation loss values for the Pythia 410M model on different adaptation datasets.

Adaptation	Dataset	Samsum			German Wiki			Bookcorpus2 Val			Bookcorpus2 Train			Github Val			Enron Val		
		$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$
Prefix		2.486	2.966	7.227	2.957	3.345	9.669	3.249	3.583	4.702	3.284	3.665	4.792	2.139	2.760	8.701	2.990	3.835	4.869
LoRA		2.403	2.830	7.176	2.880	3.276	8.365	3.125	3.454	3.219	3.119	3.490	3.333	1.698	2.288	7.484	2.588	2.708	4.170
Full fine-tune		2.415	2.690	7.867	2.892	3.084	10.101	3.104	3.577	3.506	3.133	3.616	3.153	1.851	2.768	8.616	2.845	3.715	5.681
Head fine-tune		2.481	2.813	8.382	2.877	3.122	10.567	3.123	3.428	3.733	3.118	3.460	4.032	1.721	1.952	7.905	2.590	2.753	6.037
Average		2.446	2.825	7.663	2.901	3.207	9.676	3.150	3.511	3.790	3.163	3.558	3.827	1.852	2.442	8.176	2.753	3.275	5.189

Table 23: Validation loss values for the Pythia 160M model on different adaptation datasets.

Adaptation	Dataset	Samsum			German Wiki			Bookcorpus2 Val			Bookcorpus2 Train			Github Val			Enron Val		
		$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$
Prefix		3.011	3.475	3.436	3.715	3.742	4.448	3.608	3.598	3.808	3.641	3.641	3.865	2.571	2.488	3.138	3.407	3.389	3.735
LoRA		2.702	3.038	3.180	3.458	3.459	3.578	3.396	3.420	3.537	3.400	3.423	3.690	2.020	2.050	2.444	3.003	3.023	3.119
Full fine-tune		2.486	6.803	3.062	3.396	3.624	4.284	3.396	3.562	3.422	3.402	3.588	3.739	2.025	2.263	2.855	3.083	3.154	3.382
Head fine-tune		2.862	2.883	3.425	3.418	3.445	4.048	3.402	3.417	3.694	3.432	3.599	3.801	2.111	2.212	2.947	3.091	3.021	3.668
Average		2.765	4.050	3.276	3.497	3.567	4.089	3.450	3.499	3.615	3.469	3.563	3.774	2.182	2.253	2.846	3.146	3.147	3.476

Table 24: Validation loss values for the Pythia 70M model on different adaptation datasets.

Adaptation	Dataset	Samsum			German Wiki			Bookcorpus2 Val			Bookcorpus2 Train			Github Val			Enron Val		
		$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$
Prefix		3.451	3.348	3.956	4.243	4.167	4.761	3.970	3.954	4.144	4.017	3.986	4.191	2.902	2.757	3.064	3.845	3.787	4.121
LoRA		3.071	3.524	3.450	4.024	4.007	4.141	3.757	3.735	3.862	3.717	3.744	3.963	2.322	2.357	2.606	3.424	3.448	3.580
Full fine-tune		3.107	3.059	3.828	3.912	4.138	4.639	3.698	3.906	4.073	3.707	3.940	4.090	2.402	2.651	3.074	3.420	3.587	3.792
Head fine-tune		3.108	3.336	4.488	3.977	4.070	4.148	3.719	3.745	3.891	3.745	3.968	3.862	2.412	2.715	2.940	3.514	3.727	4.307
Average		3.184	3.267	3.930	4.039	4.095	4.422	3.781	3.835	3.993	3.797	3.909	4.027	2.509	2.620	2.921	3.551	3.637	3.950

Table 25: Validation loss values for the Gpt Neo 1.3B model on different adaptation datasets.

Adaptation	Dataset	Samsum			German Wiki			Bookcorpus2 Val			Bookcorpus2 Train			Github Val			Enron Val		
		$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$
Prefix		4.154	11.172	12.590	3.306	12.510	13.110	5.016	11.610	12.862	4.590	12.119	12.848	2.889	11.377	11.868	4.133	12.400	12.231
LoRA		2.723	2.407	2.724	2.450	2.409	2.505	3.062	3.042	3.062	3.050	3.033	3.050	1.247	2.913	11.451	2.156	2.153	2.156
Full fine-tune		2.494	2.630	3.578	2.568	3.101	4.375	3.302	3.509	4.281	3.311	3.560	4.324	2.146	8.471	2.471	2.344	2.475	2.700
Head fine-tune		2.713	2.558	2.999	2.447	2.617	2.877	3.060	6.326	3.568	3.052	3.312	3.569	1.325	1.427	1.546	2.240	2.292	2.367
Average		3.021	4.692	5.473	2.693	5.159	5.717	3.610	6.121	5.943	3.501	5.506	5.948	1.902	6.047	6.834	2.718	4.830	4.878

Table 26: Validation loss values for the Gpt Neo 125M model on different adaptation datasets.

Adaptation	Dataset	Samsum			German Wiki			Bookcorpus2 Val			Bookcorpus2 Train			Github Val			Enron Val		
		$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$	$\epsilon = \infty$	$\epsilon = 8$	$\epsilon = 0.1$
Prefix		4.891	14.114	14.174	5.640	20.577	20.623	6.251	14.268	14.337	7.370	14.299	14.401	5.117	13.307	13.368	6.308	14.242	14.242
LoRA		2.694	3.070	3.073	3.243	3.244	3.244	3.491	3.491	3.491	3.504	3.492	3.492	1.605	1.595	1.595	2.766	2.757	2.757
Full fine-tune		4.716	3.252	5.524	5.195	3.244	4.492	5.551	3.494	4.398	6.623	4.728	6.499	4.133	2.859	5.329	4.854	3.483	4.663
Head fine-tune		3.178	2.867	3.512	3.176	3.500	3.641	3.472	3.773	4.255	4.304	3.908	4.280	3.093	1.928	2.194	4.064	2.955	3.020
Average		3.870	5.826	6.571	4.314	7.641	8.000	4.691	6.256	6.620	5.450	6.607	7.168	3.487	4.922	5.622	4.498	5.859	6.170

Subset	GitHub	BookCorpus2	Enron	ArXiv	CC	EuroParl	FreeLaw	USPTO	Wikipedia
Memorized Samples	192	3	18	2	8	0	7	4	2

Table 27: Set of memorized samples identified from the subsets of the Pile dataset.

## 807 H RMIA Hyperparameters

808 We focus on the importance of  $\gamma$ , as  $\alpha$  has a much more limited effect, and we set it to 0. Figure [13](#)  
809 shows the importance of  $\gamma$  and suggests that  $\gamma = 1$  is often the best choice. We omit it for simplicity,  
810 but a similar trend can be observed for the other settings.

## 811 I Broader Impact

812 Recognizing a potential underestimation of privacy risks in adapted LLMs due to insufficient empirical  
813 analysis of the combined effects of pretraining and adaptation, we conduct a rigorous benchmark. Our  
814 work offers impact by providing the community with clear guidance on privacy-preserving strategies,  
815 suitable adaptation techniques, thus contributing in more privacy-aware adapting LLMs.



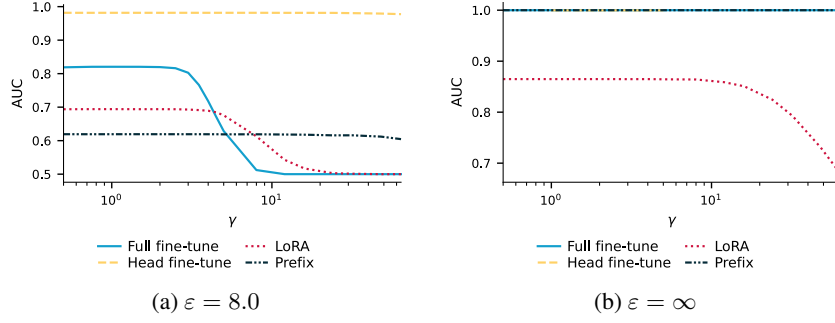


Figure 13:  $\gamma = 1$  is a strong baseline. We present the AUC using RMIA with different types of values of  $\gamma$  after adapting Pythia 1B on SAMSum. The evaluation was done for  $\varepsilon = \{8, \infty\}$ .

## J Limitations

This work focuses solely on auditing the private adaptations and leakage from pretraining data after adaptations. However, as we show, for holistic privacy auditing under the pretrain-adapt paradigm, we need ways to audit all process stages (jointly). We also focus only on a subset of models, particularly leaving out state-of-the-art closed models, such as GPT4, given that they cannot easily be adapted with DP as of the current API specification.