Benchmarking Empirical Privacy Protection for Adaptations of Large Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent work has applied differential privacy (DP) to adapt large language models (LLMs) for sensitive applications, offering theoretical guarantees. However, its practical effectiveness remains unclear, partly due to LLM pretraining, where overlaps and interdependencies with adaptation data can undermine privacy despite DP efforts. To analyze this issue in practice, we investigate privacy risks under DP adaptations in LLMs using state-of-the-art attacks such as *robust membership inference* and *canary data extraction*. We benchmark these risks by systematically varying the adaptation data distribution, from exact overlaps with pretraining data, through in-distribution (IID) cases, to entirely out-of-distribution (OOD) examples. Additionally, we evaluate how different adaptation methods and different privacy regimes impact the vulnerability. Our results show that distribution shifts strongly influence privacy vulnerability: the closer the adaptation data is to the pretraining distribution, the higher the practical privacy risk at the same theoretical guarantee, even without direct data overlap. We find that parameter-efficient fine-tuning methods, such as LoRA, achieve the highest empirical privacy protection for OOD data. Our benchmark identifies key factors for achieving practical privacy in DP LLM adaptation, providing actionable insights for deploying customized models in sensitive settings. Looking forward, we propose a structured framework for *holistic* privacy assessment beyond adaptation privacy, to identify and evaluate risks across LLMs' full pretrain-adapt pipeline.

1 Introduction

The use of *pretrained* large language models (LLMs) for sensitive downstream tasks, such as medical decision making, has grown rapidly [25, 12, 49]. To offer protection for the private data used to *adapt* the LLMs to these sensitive tasks, differential privacy (DP) [16, 17] has emerged as a gold standard [53, 54, 30, 13, 33]. However, adapting a pretrained LLM with DP may not always provide the anticipated privacy protections [48]. The challenge arises from potential overlap or complex interdependencies between data used to pretrain the LLMs and the adaptation dataset. The problem is exacerbated by the fact that for most LLMs, their pretraining datasets are not disclosed [35, 39, 46], rendering a structured reasoning of the interdependencies with the private adaptation data impossible.

While prior work has investigated privacy risks stemming from LLM pretraining [10, 9], post-hoc leakage in non-private adaptations [58], or auditing DP adaptations via synthetic canaries [36], we still lack a structured understanding of the *empirical privacy risks* of DP adaptations. This is a critical gap. Without a clear understanding of the practical risks, LLM practitioners are left with little guidance on how to privately apply LLMs in privacy-sensitive settings, including critical questions like: which adaptation method to use, what pretrained model is best given the private adaptation data distribution, and what privacy levels will be protective enough.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

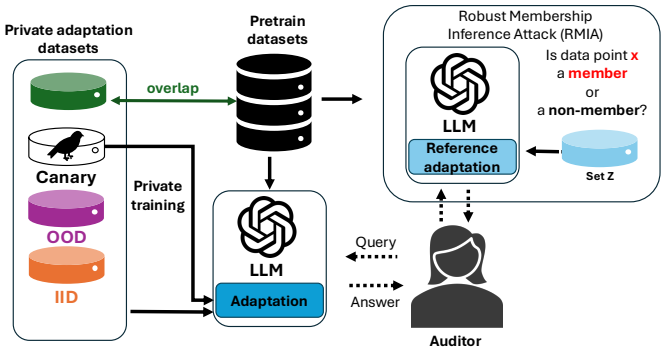


Figure 1: **Setup for Privacy Auditing of DP-LLM Adaptations.** We perform our audits based on the privately adapted LLM's output, either by using RMIA [8] as the strongest state-of-the-art membership inference attack, or by relying on data extraction attacks. For the latter, we include *canary* data into the adaptation set and measure its exposure.

To close this gap, we conduct a comprehensive benchmark evaluation that sheds light on the empirical leakage introduced by DP adaptations. We evaluate a wide range of private adaptation strategies, including full and last-layer DP fine-tuning [30], parameter-efficient fine-tuning (PEFT) methods such as DP-LoRA [21, 54], DP-Prefix Tuning [31], as well as DP prompting schemes [13]. To assess leakage, we focus on the *Robust Membership Inference Attack* (RMIA) [56], which represents the strongest state-of-the-art threat model for auditing LLM privacy, and complement this with *data extraction attacks* [47, 7, 6] to evaluate more severe forms of information leakage. A general overview of privacy auditing for adapted LLMs is provided in Figure 1.

We systematically analyze a spectrum of possible distributions for the adaptation data with respect to the pretraining data—ranging from data perfectly overlapping with the pretraining data, over IID scenarios, to entirely OOD examples—to understand the possible privacy implications for all setups. Our benchmark spans *six* datasets drawn from diverse domains, *four* adaptation methods, *and* six pretrained LLMs of different sizes and architectures, enabling comprehensive comparisons across setups. We further analyze a broad spectrum of privacy regimes from no privacy to high privacy, enabling structured reasoning about the resulting risks. Our study is guided by a central question: *What are the empirical privacy risks for the adaptation data that result from DP adaptations?*

Looking ahead, we highlight the need to jointly audit privacy risks from pretraining and adaptation and their interplay, as LLMs may leak information from either stage. To address this, we propose a new structured framework for holistic privacy assessment across the full pretrain-adapt pipeline. It defines four key audit stages: (1) pretraining, (2) adaptation, (3) their joint interaction, and (4) post-adaptation auditing of pretraining. To formally ground these audits and make them instantiatable, we redefine each stage's membership inference game [52, 23]. We hope this formalization and our practical insights from the benchmark will guide researchers in developing future assessments and help practitioners deploy customized LLMs responsibly in sensitive domains.

2 Background and Related Work

Differential Privacy. The mathematical framework of DP [16] formalizes the intuition that privacy guarantees can be obtained when a randomized mechanism $\mathcal{M}$ executed on two neighboring datasets D, D' that differ in only one data point, yields roughly the same result, *i.e.,*

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(D') \in S] + \delta. \tag{1}$$

The privacy parameter ε specifies how much the result can differ, and δ is the probability of failure to meet that guarantee. There are two canonical algorithms to implement DP guarantees in machine learning (ML): DPSGD (*Differentially Private Stochastic Gradient Descent*) algorithm [2], which extends standard stochastic gradient descent with clipping and noising gradients, and PATE (*Private Aggregation of Teacher Ensembles*) [37, 38], which is an inference time algorithm that privately transfers knowledge from an ensemble of teachers to a public student model.

Private Adaptations of LLMs. LLMs are pretrained on extensive amounts of public data, followed by adaptations to private downstream tasks. The existing methods for private LLM adaptations fall into two categories: (1) *private tuning methods*, such as PrivateLoRA [54] or PromptDPSGD [13], that rely on access to the LLM gradients and are based on the DPSGD algorithm, and (2) *private in-context learning (ICL) methods*, such as DP-ICL [51] or PromptPATE [13], which require only API (black-box) access to the LLM and are based on PATE. See Appendix A.1 for details.

Membership Inference Attacks. A membership inference attack (MIA) [44, 56, 43, 8] aims to determine whether a specific data point can be identified as part of a model's training set. This approach plays a crucial role in applications ranging from privacy assurance [45] to identifying protected or copyrighted content embedded in pretraining data [41]. While most MIA research has focused on supervised learning settings [8], new advancements reveal their broader relevance. Duan et al. [14] revealed a discrete-prompt-based MIA, disclosing vulnerabilities in proprietary LLMs like GPT-3, which risk leaking private information through prompt-based queries [13]. See Appendix A.2 for an in-depth discussion of the existing attacks.

Canary Exposure and Data Extraction Attacks. An alternative to membership inference attacks (MIAs) for evaluating privacy leakage in machine learning models is to measure the *exposure* of training data. Given a universe of candidates $\mathcal{U}$ and an attacker's ranking $\hat{Z}$ by likelihood of membership, the exposure of a target sample $z \in \mathcal{U}$ is defined as:

$$\textbf{exposure}(z, \hat{Z}) \;=\; \log_2 |\mathcal{U}| \;-\; \log_2\big(\mathrm{rank}(z; \hat{Z})\big). \tag{2}$$

This score is maximal when z is ranked most likely and zero when ranked least likely. In a complementary vein, *extractability* quantifies how readily a model emits a secret string when prompted. A suffix s is said to be *extractable with k tokens of context* if there exists some prefix p of length k such that, under greedy decoding, the model outputs s immediately following p. When s is sufficiently long and random, its extractability serves as a practical metric of memorization in LLMs. Further discussion appears in Appendix A.3.

Benchmarking Privacy Vulnerabilities. Zhu et al. [58] introduced *PrivAuditor*, which systematically and empirically evaluates the privacy leakage from LLM adaptations. In contrast to our work, they focus on *non-private* adaptations only. Li et al. [27] evaluated the privacy leakage of private LLMs adaptations through empirical privacy attacks, such as data extraction, MIAs, and embedding-level privacy attacks. This benchmark focuses mostly on tradeoffs between privacy and utility, highlighting the complexity of balancing them. Contrary to our work, this work does not explore the relationship between the pretraining data and the fine-tuning one. *LLM-PBE* [28] empirically evaluates privacy risks throughout the LLM lifecycle, including pretraining, fine-tuning, and querying. Zhou et al. [57] investigated potential data leakage across widely used software engineering benchmarks.

3 Experimental Setup

We begin by detailing the setup used for our benchmark. Further details are presented in Appendix B.

Models and Pretraining Data. Our work primarily focuses on the Pythia family of models trained on the Pile dataset [18], and the GPT-Neo family [4]. To benchmark the effects over various model sizes, we use Pythia 1.4B, Pythia 1B, Pythia 410M, Pythia 160M, Pythia 70M, GPT Neo 1.3B, and GPT Neo 125 M. The Pile dataset [18] is an 800GB collection of diverse English-language datasets, including text from sources such as books, academic papers, or source code repositories. In all cases where a specific model is not explicitly mentioned, we use Pythia 1B as the default model.

Adaptation Datasets. We categorize the datasets used in our experiments into **in-distribution (IID)** and **out-of-distribution (OOD)**, depending on their relationship to the pretraining data. IID datasets come from the same distribution as the pretraining data, and we identify two cases: one with a full overlap between pretraining and adaptation data, where we use data directly from the pretraining set for the adaptations, and one with no overlap, where the data is sourced from the corresponding validation set from the pretraining distribution. We focus on the following Pile subsets for the IID datasets: BookCorpus2, GitHub, and Enron Emails [24]. In contrast, OOD datasets are derived from a different distribution and do not overlap with pretraining data. Thereby, we choose SAMSum [19], and GermanWiki [1]. We elaborate more in Appendix B.1.

Adaptation Methods. We evaluate different types of adaptations, including fine-tuning of all model parameters [30], or the last layer (*i.e.,* the head) and PEFT methods, such as LoRA [21, 54] and Prefix

Tuning [31, 13]. Considering a Pythia 1B model, we train 1B parameters for Full Fine-Tuning, 1M for LoRA, 130M for Prefix Tuning, and 100M for last-layer (Head) Fine-Tuning. Since membership inference success is highly dependent on the train-test gap, for a fair comparison of the privacy leakage, we ensure similar evaluation perplexities, in particular, similar validation loss values at the end of the adaptation's training for specific datasets across adaptation methods, see Appendix B.2.

Membership Inference. For membership inference, we rely on the strongest state-of-the-art attack, namely RMIA (Robust Membership Inference Attack) [56]. We use its offline version because it is computationally effective and does not require training customized reference models for each targeted sample (as in the online version of the attack). We also leverage a single reference model for our experiments, as the authors show strong MIA performance even with a single reference model. We consider different types of reference models. Unless explicitly stated, we focus on using a "shadow" model (adaptation), in our case Pythia 1B, which is trained in the same way as the target model, but on a different split of the same fine-tuning data. We also evaluate the *Reference* method [7], which calibrates the target model's loss using a reference model, and compare against Min-K% as a reference-less baseline attack. As with RMIA, we report the best AUC from a grid search over Min-K%'s parameter K. See Appendix B.4 for a detailed description of the setup.

Canary Exposure and Data Extraction Attacks. To evaluate memorization, we insert adversarial canaries into a small portion of the adaptation data and estimate their exposure using two approximation methods: sampling and distribution modeling. Both approaches perform similarly when using 256 non-member canaries, and we adopt sampling for efficiency. Moreover, when considering k-extractable memorization, we set $k = 10$ tokens. A detailed description of the data extraction setup is provided in Appendix B.5.

4 Benchmark design and experiments

To address our benchmark's central question—*What are the empirical privacy risks to adaptation data under DP adaptations?*—we break it down into five concrete research questions.

4.1 RQ1: How does the relationship (overlapping, IID, OOD) between adaptation and pretraining datasets impact data privacy?

Motivation. The pretrain-adapt paradigm uses LLMs pretrained on large public datasets, which are then adapted to smaller, often sensitive, private datasets using DP methods. While DP offers formal guarantees, its practical effectiveness under the pretrain-adapt paradigm remains unclear—particularly how the relationship and interplay between adaptation and pretraining data (*e.g.,* overlapping, IID, or OOD) influences actual privacy leakage.

Summary of Findings. Our results show that (1) privacy risks increase when the adaptation data distribution is closer to the pretraining data, even if there is no direct overlap. (2) Surprisingly, IID data from the pretraining validation set leaks as much as directly overlapping data, underscoring distributional closeness as the main driver of risk.

Detailed Results. We present our main results in Table 1 and Table 2. We focus our discussion on Pythia-1B, and further expand it for the other models in Appendix C.1. They show that the average AUC is generally higher in IID settings than OOD in all attacks and adaptations. For instance, looking at *RMIA (shadow)* using $\varepsilon = 8$, we observe that the average AUC is between 0.7 and 0.9 in the IID setting, while it is between 0.63 and 0.87 for the OOD setting. More detailed analyses for different attack setups and more privacy regimes are depicted in Appendix C.1. We also identify distributional closeness as a key risk factor, as overlapping data leaks similarly to IID. Moreover, our results indicate that under both a strong attack and in more practical scenarios, moderate privacy regimes (*e.g.,* $\varepsilon = 8$) still present a real threat of privacy leakage from IID. On the other hand, under this regime, privacy leakage from the OOD is mostly observed with a strong attack. Moreover, in Appendix C.4, Figure 8 shows over the training epochs the Overlap (Train) and IID data (Val) privacy leakage, and further highlights a similar privacy leakage between Overlap and IID data across the whole training run. We also analyze the impact of subset characteristics on privacy leakage in Appendix C.3, and we discover that the pretraining dataset size and complexity influence the privacy leakage in the training datasets. We observe that privacy leakage increases with both the size and complexity of the subsets. Larger datasets produce more IID results than smaller subsets, further validating our findings.

Table 1: **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 RMIA MIAs for the Pythia 1B model adapted on different datasets with $\varepsilon \in \{0.1, 8, \infty\}$.

MIA	Adaptation \ Dataset	SAMSum			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
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

Table 2: **Membership Inference for in-distribution (IID) Adaptations** using the setup from Table 1.

MIA	Adaptation \ Dataset	Bookcorpus2 Val			Bookcorpus2 Train			Github Val			Enron Val			Average		
		$\varepsilon=\infty$	$\varepsilon=8$	$\varepsilon=0.1$	$\varepsilon=\infty$	$\varepsilon=8$	$\varepsilon=0.1$	$\varepsilon=\infty$	$\varepsilon=8$	$\varepsilon=0.1$	$\varepsilon=\infty$	$\varepsilon=8$	$\varepsilon=0.1$	$\varepsilon=\infty$	$\varepsilon=8$	$\varepsilon=0.1$
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
	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
	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
	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
	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
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
	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
	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
	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
	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

4.2 RQ2: Which DP adaptation method is the most protective?

Motivation. It is known that the type of adaptation has a significant impact on the utility of the final model [58]. However, different adaptations might also offer disparate empirical protection at the same formal privacy guarantee, motivating our empirical comparison.

Summary of Findings. While LoRA provides much better empirical privacy protection in non-private settings compared to other adaptations, the differences become more subtle under the DP regime. Despite this, LoRA consistently achieves a relatively low AUC, whereas the other adaptations show varying trends depending on the dataset or privacy budget.

Detailed Results. Specifically, as shown in Table 1 for OOD datasets with $\varepsilon = 8$, the most vulnerable adaptations are Full and Head Fine-Tune. On the other hand, for IID data, the strongest protection provides Head Fine-Tune, which is marginally better LoRA. With stronger privacy guarantees, LoRA is the most private for OOD datasets with an AUC score of 0.58, thus slightly better than Full Fine-Tune. On the other hand, while adapting to the IID dataset, LoRA outperforms other adaptations. Notably, Full Fine-Tune and Head Fine-Tune show much lower privacy protection in these settings.

4.3 RQ3: Are the same adaptations robust against data extraction?

Motivation. Data extraction attacks are even more severe than MIAs. Therefore, it is crucial to evaluate the protectiveness of DP adaptations against this stronger threat.

Summary of Findings. We find that Prefix Tuning is the most vulnerable adaptation method in this setting. On the other hand, LoRA and Head Fine-Tune in both cases, with and without DP guarantees exhibit resistance against data extraction.

Detailed Results. We report detailed results in Appendix C.2. In particular, Table 17 and Table 18 show that for $\varepsilon = 0.1$ the exposure is around 1.44, therefore, close to random guessing. We also noticed a limited influence on the choice of the canary prefix type. Moreover, the adversarial prefix is the main source of privacy leaks, with the interaction between the prefix and the individual sample playing a smaller role, see Figure 9 in Appendix C.5.

4.4 RQ4: How important is the attacker's knowledge of the pretrained model?

Motivation. The attacker's knowledge of the pretrained model plays a crucial role in the success of MIAs, as it enables them to select more relevant reference models and non-member data for training,

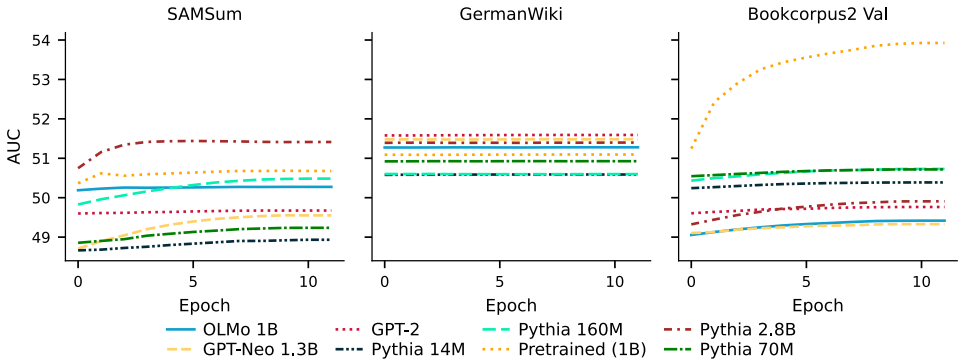


Figure 2: **IID data is more susceptible to leakage using the pretrained base model than OOD data.** We compare the effectiveness of performing RMIA on fully fine-tuned Pythia 1B with $\varepsilon = 8$ with different pretrained models as reference models.

which is one of the main challenges of MIAs [50, 8]. We investigate various setups, including an attacker who has access to a shadow model from the same pretraining distribution as the adapted LLM, a similar model, and no access to external models. This helps us characterize the landscape of potential real-world risks and setups.

Summary of Findings. MIAs' performance highly depends on the attacker's knowledge of the target model and pretraining data. In particular, RMIA performs best when a shadow model shares architecture, initialization weights, and training data distribution. Meanwhile, MIAs' effectiveness rapidly deteriorates as shadow models are trained on different distributions or architectures. Particularly, we observe that when a shadow model trained on the same distribution of the target model is available, using the pretrained model is the second-best choice, followed by models of the same family and similar size.

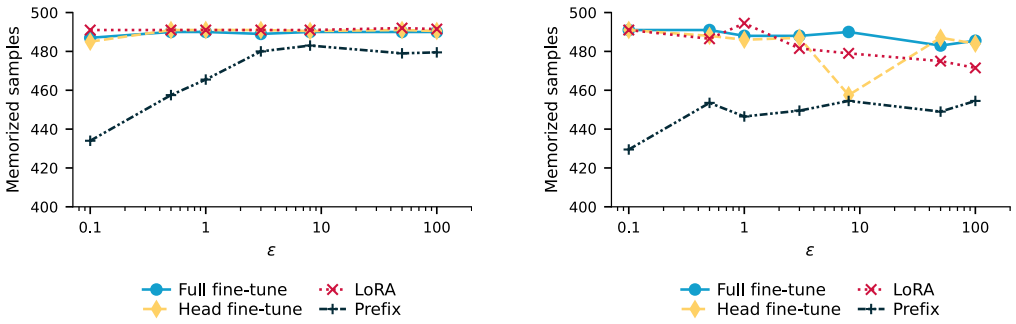
Detailed Results. To simulate attackers with various background knowledge, in this setting, we also consider other "shadow" models: Pythia 14M, Pythia 160M, Pythia 1B, Pythia 2.8B [3], GPT-neox [4], OLMo-1B [20], and GPT-2 [40]. The MIA performance is close to random for private adaptations with $\varepsilon = 8$. Furthermore, as shown in Figure 2, while the MIA's performance for Pythia 1B is higher on IID data, the choice of reference model has little effect when attacking models adapted on OOD data, even with architectural differences between the model and the reference model *i.e.,* GPT-Neo 1.3B and OLMo 1B. Moreover, as in the other case, Figure 11 (in Appendix D) shows that the privacy leakage is similar between IID and the corresponding overlapping data. We show further experiments in Appendix D.

4.5 RQ5: How does adaptation change the pretraining dataset vulnerability?

Motivation. DP adaptations only guarantee protection for the adaptation dataset. Yet, adapting the model to other data, while introducing noise, can also affect the pretraining leakage. This is an important aspect to study, as also pretraining data can be private [48], *e.g.,* private conversations with ChatGPT used to improve the models, or emails used to pretrain Gemini. Therefore, we also empirically investigate how adapting pretrained LLMs affects the leakage of pretraining data.

Summary of Findings. Our findings show that the choice of adaptation method impacts the privacy of pretraining data. Specifically, our evaluation shows that Prefix Tuning reduces the leakage of memorized pretraining data from adapted language models, especially in high-privacy settings. However, for the other adaptations, this effect is negligible, and the adapted model retain most of the pretraining memorization.

Detailed Results. We evaluate the effect of OOD and IID adaptation data on the leakage of memorized pretraining data from the adapted LLM. Specifically, as we show in Figure 3, Prefix Tuning significantly reduces leakage, particularly in high-privacy regimes. For the other adaptation methods, the number of memorized samples often remains above 460 samples. For Prefix Tuning, the number of memorized samples is often lower than 460 and goes down to around 430 with $\varepsilon = 0.1$, thus suggesting that adaptation partially mitigates the pretraining memorization.



(a) Bookcorpus2 Val

(b) SAMSum

Figure 3: **Fewer memorized samples after prefix tuning.** There are fewer verbatim generations of training samples after the prefix tuning, especially for small ε values. We present the number of memorized samples from the Pile that remain memorized after adapting Pythia 1B on Bookcorpus2 val and SAMSum datasets. The evaluation was done for $\varepsilon = \{0.1, 1, 3, 8, 50, 100, \infty\}$. We present the x-axis using a log scale.

5 Discussion of our Results

Our findings reveal a complex interplay between pretraining and adaptation data. This significantly affects the privacy risks under DP adaptations. Below, we discuss the implications of these findings when adapting pretrained LLMs to sensitive domains using DP.

Disparate Leakage Based on Distribution. Our results demonstrate that the distributional closeness between pretraining and adaptation data is a key factor influencing empirical privacy leakage under DP. Adaptations using IID data—data from the same distribution but not seen during pretraining—consistently showed the highest vulnerability. This presents a fundamental trade-off: while adapting a model already pretrained on similar data is often beneficial for utility, it simultaneously increases privacy risk.

Disparate Leakage Based on Adaptation Method. We also observe that not all DP adaptation methods offer equal protection, even when enforcing the same formal level guarantee, expressed in the same ε. This aligns with earlier findings in the non-private regime, where privacy-utility trade-offs differ across methods [58]. In our experiments, LoRA appeared most consistently robust against privacy attacks, while Prefix Tuning showed the least vulnerability to extraction attacks. These differences are highly relevant for practice: in addition to choosing methods that optimize downstream performance, practitioners should also consider empirical privacy leakage. The attacks we use in this paper offer a way to assess and understand such risks under realistic conditions.

Choosing a Privacy Regime. We find that in moderate privacy regimes, *e.g.,* $\varepsilon = 8$, sensitive adaptation data still experiences significant practical vulnerability against both MIAs and data extraction attacks. This highlights the necessity to perform private LLM adaptations in the high-privacy regime, *i.e.,* with low ε to achieve practical protection.

Reliance on Accurate Shadow Model. We show that attackers gain a substantial advantage when they have access to the original pretrained LLM used during adaptation. Shadow models instantiated with the same pretrained model as the adapted LLM's base consistently achieved higher attack success. This is especially concerning given the rise of adapting publicly available LLMs, which makes strong shadow models easily accessible to adversaries. These findings further underscore the need for stringent privacy settings in DP adaptations.

Towards a Holistic Privacy Auditing for LLMs Our results suggest that privacy assessments should not treat pretraining and adaptation in isolation. The strong interdependence between these stages demands holistic analysis. Motivated by this insight, we introduce a structured framework in the next section that formalizes how privacy assessments and audits under the pretrain-adapt paradigm should be conducted. We hope this framework encourages the development of privacy assessment methods that match the complexity of modern private LLM pipelines.

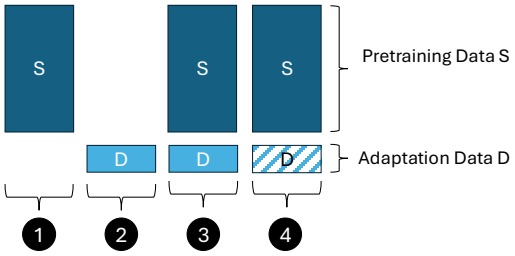


Figure 4: **Stages of Auditing.** We analyze four stages of auditing: ❶ Audit Pretraining, ❷ Audit Adaptations, ❸ Joint Auditing of Pretraining and Adaptations, ❹ Post-Adaptation Auditing of the Pretraining.

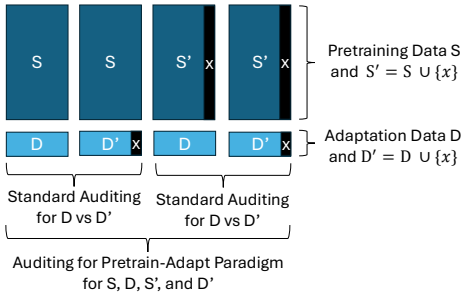


Figure 5: **Setup for Joint Adaptation auditing (3).** We consider different datasets for pretraining and adaptation, distinguishing it from standard ML privacy auditing [34, 55] by considering pretraining data.

6 Towards Holistic Privacy Audits under the Pretrain-Adapt Learning Paradigm

6.1 From Stages to Adversary Game under Pretrain-Adapt Privacy Auditing

While our understanding of empirical privacy risks has grown, we recognize the need to go further and adopt more nuanced approaches to tackle privacy risks posed during the adaptation of LLMs. Therefore, we formalize a framework to assess privacy risks holistically for LLMs and their pretrain-adapt paradigm. In total, we identify four different stages of auditing that need to be considered (see Figure 4) under the pretrain-adapt paradigm, namely (1) audit pretraining, (2) audit adaptations, (3) joint audit of pretraining and adaptations, and (4) post-adaptation auditing of the pretraining, as shown in Figure 4. Based on them, we formalize how to instantiate these audits and contrast them with standard privacy auditing. Privacy audits can be modeled as an *adversarial game* $\mathcal{G}$ [52, 23] where the main task is to guess if a given data point x was in a model's training set or not. This game can, therefore, also be referred to as the *membership inference game*. We define the adversarial game $\mathcal{G}$ analogous to the one for standard ML, yet take two datasets, S the pretraining data, and D the adaptation data into account. Additionally, we denote the pretraining procedure by T and the adaptation procedure by T'. We mark the deviations to the original game in blue.

1. The challenger samples $a \xleftarrow{R} \{0,1\}$ and $b \xleftarrow{R} \{0,1\}$ (where a and b are binary variables)
2. The challenger trains a model $\theta \xleftarrow{T} \tilde{S}, \theta_0$, where $\tilde{S} = S$ if $a = 0$, otherwise $\tilde{S} = S \cup \{x\}$
3. The challenger adapts θ such that $\theta' \xleftarrow{T'} \tilde{D}$, where $\tilde{D} = D$ if $b = 0$, otherwise $\tilde{D} = D \cup \{x\}$
4. The challenger sends θ' to the attacker
5. The attacker guesses $\hat{a}, \hat{b} \leftarrow \mathcal{A}(\theta, \theta', x)$

Whether the attacker has to guess both $\hat{a}, \hat{b}$ and what background knowledge they have, *i.e.,* whether they get access to both θ and θ' depends on the auditing stage. We detail the attacker's background knowledge and guesses—formulated as hypotheses with a null hypothesis H_0 and an alternative hypothesis H_A—for the four auditing stages from our taxonomy.

(1) Auditing pretraining resembles standard ML auditing, targeting privacy leakage from pretrained models. Differences arise from larger datasets and models, limiting both DP protection efficacy [10] and applicability of auditing techniques like MIA [15]. In this setting, the challenger releases the pretrained model θ to the attacker. The attacker's goal is to correctly guess whether x was in the pretraining data S. Their guesses $\hat{a}$, are over the random variable a.

$$H_0 : a = 0 \qquad H_A : a = 1$$

(2) Auditing adaptation a new pretrain-adapt paradigm aspect, detects adaptation dataset leakage from adapted LLMs. The key differentiating factor of privacy audits in standard ML is using a

pretrained model that the adaptations are trained on instead of a random initialization. We assume the same pretrained model is used for all the considered adaptations in an adaptation audit. In this setting, the challenger releases only the adapted model θ' to the attacker. The attacker does not know whether $x \in S$ or not and considers only the adaptation. Their guesses $\hat{b}$, are, hence, over the random variable b.

$$H_0 : b = 0 \qquad H_A : b = 1$$

(3) Joint auditing evaluates combined leakage from both pretraining and adaptation datasets in the adapted LLM. Typical privacy preservation involves non-DP-trained LLMs with DP-trained adaptations. In this setting, the challenger releases both the pretrained model θ and the adapted θ' to the attacker. Depending on the attacker's background knowledge, we consider three possible cases

The attacker knows that $x \notin S$ and guesses b.		The attacker knows that $x \in S$ and guesses b.		The attacker knows that the target sample x is either in both (pretraining and adaptation sets) or neither of them and guesses (a, b).	
$H_0 : (a, b) = (0, 0)$	$H_A : (a, b) = (0, 1)$	$H_0 : (a, b) = (1, 0)$	$H_A : (a, b) = (1, 1)$	$H_0 : (a, b) = (0, 0)$	$H_A : (a, b) = (1, 1)$

(4) Post-Adaptation Auditing evaluates how the (private) adaptations influence the potential protection of the data points used for pretraining, which is usually conducted without any formal guarantees. Changes to the model behavior induced through adaptations or noise added during their training might influence the effective exposure of pretraining data from model predictions. In this setting, the challenger releases both the pretrained θ and the adapted θ'. It is known that the target sample x is not in D and the attacker guesses a.

$$H_0 : (a, b) = (0, 0) \qquad H_A : (a, b) = (1, 0)$$

In essence, auditing pretraining considers only the pretraining itself. Similarly, auditing the adaptations considers the adaptations themselves. On the other hand, the joint adaptation reasons about both pretraining and adaptation sets. Finally, the post-adaptation auditing is only for the pretraining set, but the applied adaptation influences the auditing.

6.2 Practical Application of Holistic Audits

Our new perspective on the pretrain-adapt paradigm gives both practitioners and researchers clearer insights into each threat model's risks. Formalizing the auditing setup supports systematic reasoning about privacy risks, thus clarifying the guarantees that different methods need to provide. Therefore, our formalization allows for creating a unified interface for measuring privacy leakage, regardless of whether its source is pretraining or adaptation data. Moreover, our work demonstrates that looking at pretraining and adaptation components separately can lead to a false impression of privacy. The connection between these stages affects privacy leakage, which makes comprehensive auditing essential within pretrain-adapt paradigm. We believe that developing and sharing tools that support all privacy assessment stages, from threat modeling and risk quantification to mitigation, will empower the research community to more effectively define risks and allow for the reduction of privacy risks in practice.

7 Conclusions

In this work, we benchmark the practical privacy risks that arise under DP adaptations of LLMs within the pretrain-adapt paradigm. Our comprehensive empirical analysis confirms the theoretical concern that pretraining significantly amplifies the privacy risks associated with the *adaptation data*. We find that the closeness of adaptation and pretraining data distributions plays a critical role: even in the absence of overlap, higher distributional similarity results in increased privacy leakage. Additionally, we observe that the choice of adaptation method impacts privacy leakage, with PEFT methods, such as LoRA, offering significantly lower privacy risks while maintaining strong utility. Furthermore, we show Prefix Tuning can reduce the leakage of pretraining data, likely due to the added input noise during private adaptation. Our findings highlight the need for stringent DP constraints (*e.g.,* $\varepsilon < 0.1$) to mitigate privacy risks in LLM adaptations effectively. It also motivates the need for holistic privacy assessments under the pretrain-adapt paradigm and takes the first step towards it by formalizing such an assessment over the different stages. This work lays a foundational framework for future research efforts aimed at safeguarding privacy within the pretrain-adapt paradigm.

References

[1] Sujet-finance-instruct-177k dataset. URL https://huggingface.co/datasets/Cohere/wikipedia-22-12-de-embeddings.

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

[3] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.

[4] Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL https://doi.org/10.5281/zenodo.5297715. If you use this software, please cite it using these metadata.

[5] Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. Bad characters: Imperceptible nlp attacks, 2021. URL https://arxiv.org/abs/2106.09898.

[6] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, Santa Clara, CA, August 2019. USENIX Association. ISBN 978-1-939133-06-9. URL https://www.usenix.org/conference/usenixsecurity19/presentation/carlini.

[7] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, 2021. URL https://arxiv.org/abs/2012.07805.

[8] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022.

[9] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.

[10] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

[11] Ting-Yun Chang, Jesse Thomason, and Robin Jia. Do localization methods actually localize memorized data in LLMs? a tale of two benchmarks. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3190–3211, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.176. URL https://aclanthology.org/2024.naacl-long.176.

[12] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. Meditron-70b: Scaling medical pretraining for large language models, 2023. URL https://arxiv.org/abs/2311.16079.

[13] Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. Flocks of stochastic parrots: Differentially private prompt learning for large language models. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[14] Haonan Duan, Adam Dziedzic, Mohammad Yaghini, Nicolas Papernot, and Franziska Boenisch. On the privacy risk of in-context learning. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

[15] Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=av0D19pSkU.

[16] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer, 2006.

[17] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[18] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[19] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5409. URL https://www.aclweb.org/anthology/D19-5409.

[20] Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, 2024.

[21] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[22] Matthew Jagielski. A note on interpreting canary exposure. *arXiv preprint arXiv:2306.00133*, 2023.

[23] Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, and David Evans. Revisiting membership inference under realistic assumptions. *arXiv preprint arXiv:2005.10881*, 2020.

[24] Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, 2004. URL https://api.semanticscholar.org/CorpusID:265038669.

[25] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

[26] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, November 2021. URL https://aclanthology.org/2021.emnlp-main.243.

[27] Haoran Li, Dadi Guo, Donghao Li, Wei Fan, Qi Hu, Xin Liu, Chunkit Chan, Duanyi Yao, Yuan Yao, and Yangqiu Song. PrivLM-bench: A multi-level privacy evaluation benchmark for language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 54–73, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.4. URL https://aclanthology.org/2024.acl-long.4/.

[28] Qinbin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, Bo Li, Bingsheng He, and Dawn Song. Llm-pbe: Assessing data privacy in large language models, 2024. URL https://arxiv.org/abs/2408.12787.

[29] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353.

[30] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=bVuP3ltATMz.

[31] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

[32] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.8. URL https://aclanthology.org/2022.acl-short.8.

[33] Harsh Mehta, Walid Krichene, Abhradeep Guha Thakurta, Alexey Kurakin, and Ashok Cutkosky. Differentially private image classification from features. *Transactions on Machine Learning Research*, 2023.

[34] Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 1631–1648, 2023.

[35] OpenAI. Gpt-4 technical report, 2023.

[36] Ashwinee Panda, Xinyu Tang, Milad Nasr, Christopher A. Choquette-Choo, and Prateek Mittal. Privacy auditing of large language models. In *ICML 2024 Next Generation of AI Safety Workshop*, 2024. URL https://openreview.net/forum?id=6mVZUh4kkY.

[37] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian J. Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[38] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with PATE. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[39] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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

[41] Avital Shafran, Shmuel Peleg, and Yedid Hoshen. Membership inference attacks are easier on difficult problems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14820–14829, October 2021.

[42] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zWqr3MQuNs.

[43] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models, 2024. URL https://arxiv.org/abs/2310.16789.

[44] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[45] Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=f38EY21lBw.

[46] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[47] Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. Truth serum: Poisoning machine learning models to reveal their secrets. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2779–2792, 2022.

[48] Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Position: Considerations for differentially private learning with large-scale public pretraining. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 48453–48467. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/tramer24a.html.

[49] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024.

[50] Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=3eIrli0TwQ.

[51] Tong Wu, Ashwinee Panda, Jiachen T. Wang, and Prateek Mittal. Privacy-preserving in-context learning for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=x4OPJ7lHVU.

[52] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.

[53] Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Large scale private learning via low-rank reparametrization. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12208–12218. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/yu21f.html.

[54] 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. Differentially private fine-tuning of language models. In *International*

Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=Q42f0dfjECO.

[55] Santiago Zanella-Beguelin, Lukas Wutschitz, Shruti Tople, Ahmed Salem, Victor Rühle, Andrew Paverd, Mohammad Naseri, Boris Köpf, and Daniel Jones. Bayesian estimation of differential privacy. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 40624–40636. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zanella-beguelin23a.html.

[56] Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks, 2024. URL https://arxiv.org/abs/2312.03262.

[57] Xin Zhou, Martin Weyssow, Ratnadira Widyasari, Ting Zhang, Junda He, Yunbo Lyu, Jianming Chang, Beiqi Zhang, Dan Huang, and David Lo. Lessleak-bench: A first investigation of data leakage in llms across 83 software engineering benchmarks. *arXiv preprint arXiv:2502.06215*, 2025.

[58] Derui Zhu, Dingfan Chen, Xiongfei Wu, Jiahui Geng, Zhuo Li, Jens Grossklags, and Lei Ma. Privauditor: Benchmarking data protection vulnerabilities in llm adaptation techniques. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/12b18a15dcd73e1991e9959a94375fab-Paper-Datasets_and_Benchmarks_Track.pdf.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 4 explains in detail each contribution.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix J discuss the limitations of the work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our work does not include any theoretical results.

   Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide a detailed description of the experimental settings in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the code in the supplemental material, and will release the code used for all the experiments.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix B describe the setup in details, and Appendix B.3 describe the hyperparameter selection.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars due to the high computational cost of replicating our experiments. Each model adaptation at different scales and architectures requires substantial resources, and multiple runs for statistical validation would greatly increase the overall cost.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).

- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the most computationally expensive steps in Appendix B.7.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors fully comply with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the societal impact of our work in Appendix I.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our research relies on publicly available models and datasets, presenting no significant risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Our code includes no source code or binary files from external libraries. Therefore, there are no concerns regarding permissions or license inclusion. We use only open-source datasets and models, all of which are properly cited in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide a code base along with documentation. In Appendix B we provide a detailed description of the used datasets, models, adaptations and hyperparameters.

    Guidelines:

    - The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were not used for the impact of the core methodology, scientific rigor, or originality of the research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.