Benchmarking Empirical Privacy Protection for Adaptations of Large Language Models

Bartłomiej Marek*, Lorenzo Rossi*, Vincent Hanke, Xun Wang,
Michael Backes, Franziska Boenisch, Adam Dziedzic

CISPA Helmholtz Center for Information Security

ICLR International Conference On Learning Representations
ORAL

Paper GitHub

TL;DR

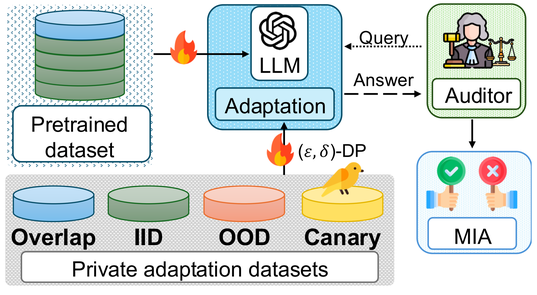
We benchmark empirical privacy in DP-adapted LLMs and show that **data closer to pretraining** leaks more, **LoRA** is the most consistent choice overall, and $\varepsilon = 8$ often remains vulnerable.

Main Contributions

This work benchmarks empirical privacy under differentially private adaptation of pretrained LLMs.

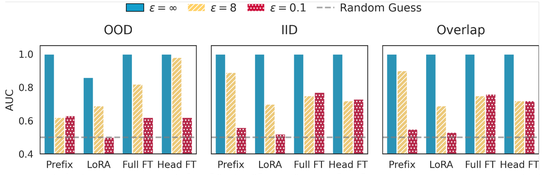
- Benchmarks privacy risks across **overlap**, **IID**, and **OOD** settings using strong membership inference and extraction attacks.
- Compares **Full FT**, **Head FT**, **LoRA**, and **Prefix Tuning** across models, privacy budgets, and attacker settings.
- Shows that data **closer to pretraining** leaks more in practice, even without exact overlap.
- Identifies **LoRA** as the most consistent privacy–utility choice overall, while showing that moderate ε can remain vulnerable in practice.
- Introduces a **holistic auditing framework** for the pretrain–adapt pipeline, covering pretraining, adaptation, and their interaction.

Overview of the Benchmark



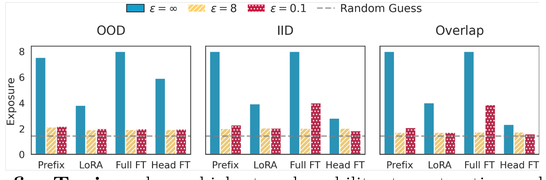
We consider the test subsets as IID data, while GermanWiki and SAMSum serve as OOD datasets that differ linguistically and structurally, respectively.

Distribution Impact on Privacy



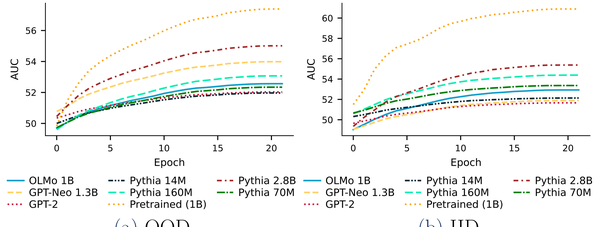
Privacy risk increases with **distributional closeness** to pretraining. **IID** data leaks nearly as much as overlapping data.

Data Extraction Resistance



Prefix Tuning shows highest vulnerability to extraction, while **LoRA** and **Head Fine-Tuning** show strongest resistance.

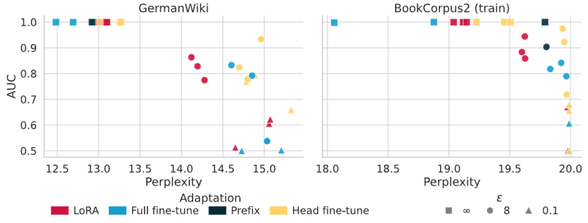
Attacker's Knowledge Is Crucial



(a) OOD (b) IID

Attack strength depends on the reference model; a shadow model works best, and the pretrained base is the strongest fallback.

Privacy vs Utility



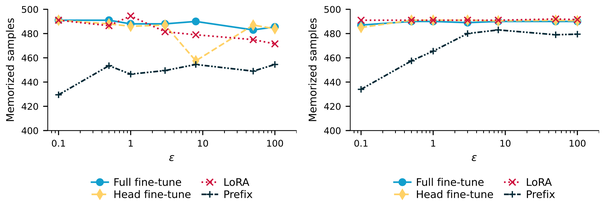
At comparable perplexity, **LoRA** usually attains lower AUC than the other adaptation methods. It therefore gives the strongest overall privacy–utility trade-off in this benchmark.

Which Adaptation is Most Protective?

Method	German Wiki	BC2 Val	BC2 Train	GitHub Val
Prefix Tuning	0.64	0.89	0.90	0.93
LoRA	**0.59**	**0.70**	**0.69**	**0.74**
Full FT	0.71	0.75	0.75	0.78
Head FT	0.76	0.72	0.72	0.80

BC2 denotes BookCorpus2. Across the shown $\varepsilon = 8$ cases, **LoRA** achieves the lowest AUC and is the clearest choice at this privacy level.

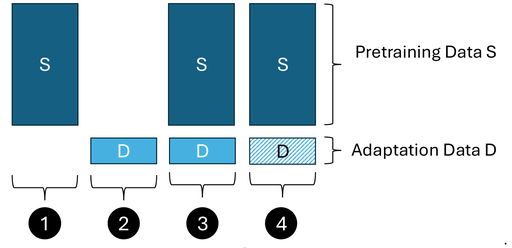
Effect on Pretraining Leakage



(a) OOD (b) IID

DP adaptation also changes the leakage of memorized pretraining data; Prefix Tuning reduces it the most, especially at high privacy levels.

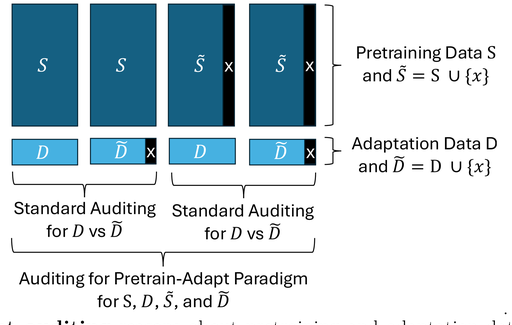
Holistic Privacy Auditing Framework



1 Audit pretraining leakage 2 Audit adaptation leakage
3 Joint audit 4 Post-adaptation audit of pretraining

We formalize **four audit stages** for the pretrain–adapt pipeline. This matters because auditing only adaptation can miss leakage from **pretraining** or from its interaction with adaptation.

Setup for Joint Adaptation Auditing



Joint auditing reasons about pretraining and adaptation datasets together with the two training stages. This is the right threat model when privacy risk may come from either stage or from their interaction.