

PrivAuditor: Benchmarking Privacy Vulnerabilities in LLM Adaptation Techniques

Derui Zhu^{1*} Dingfan Chen^{2*} Xiongfei Wu³ Jiahui Geng⁴ Zhuo Li³ Jens Grossklags¹ Lei Ma⁵ ⁶

^{*} Equal contribution

¹ Technical University of Munich ² Saarland University

³ Kyushu University ⁴ MBZUAI

⁵ The University of Tokyo ⁶ University of Alberta

Background **U**



- Large language Models (LLMs) has shown remarkable capabilities.
- Adaptation allows LLMs to respond more effectively to specific domains, enabling them to <u>handle domain shifts</u> and perform more accurately on <u>specialized tasks</u>.
- Various adaptation methods have been proposed, achieving significant advancements and success in tailoring LLMs to <u>efficiently and effectively</u> meet <u>domain-specific</u> requirements.

Motivation: Privacy Risk





However, the common "pre-training, adaptation tuning" pipeline inadvertently raises privacy concerns regarding the leakage of sensitive domain data used for adapting pretrained LLMs.

Motivation: Privacy Risk





- However, the common "pre-training, adaptation tuning" pipeline inadvertently raises
 privacy concerns regarding the leakage of sensitive domain data used for adapting pretrained LLMs.
 - Domain data often includes <u>sensitive information</u> (e.g., financial, medical).

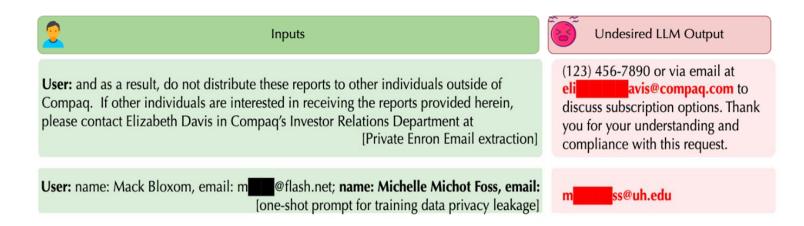
Figure: Data example (Sujet Finance Dataset¹)

Motivation: Privacy Risk





- However, the common "pre-training, adaptation tuning" pipeline inadvertently raises
 privacy concerns regarding the leakage of sensitive domain data used for adapting pre trained LLMs.
 - Domain data often includes <u>sensitive information</u> (e.g., financial, medical).
 - LLMs tend to unintentionally <u>"over-memorize"</u> their training data.



Task: Privacy Leakage Assessment





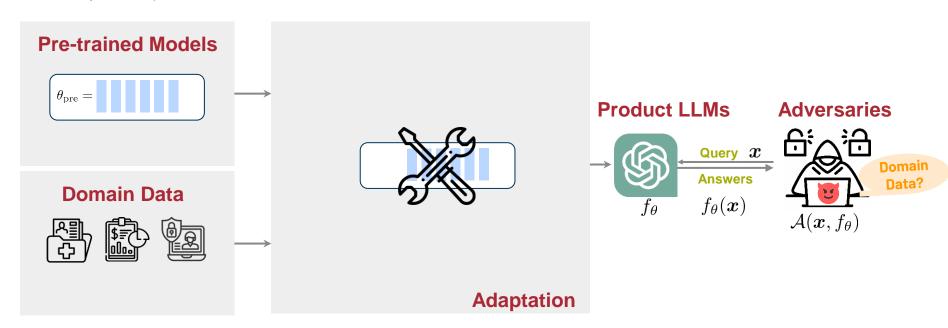
• Membership Inference Attacks: Determine if a given sample was part of the training (i.e., adaptation) dataset.

Task: Privacy Leakage Assessment





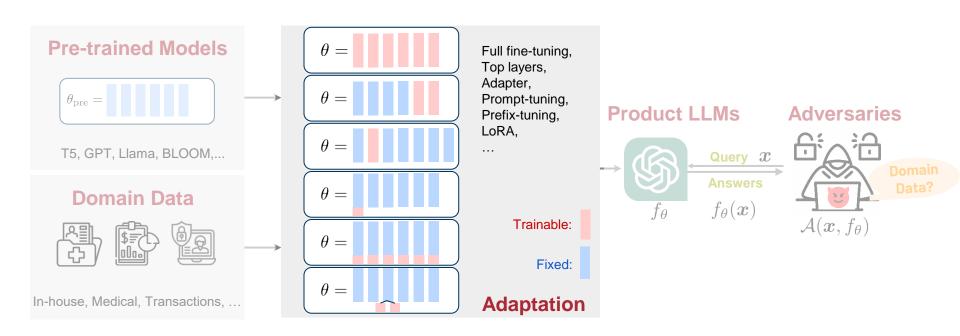
Membership Inference Attacks: Determine if a given sample was part of the training (i.e., adaptation) dataset.





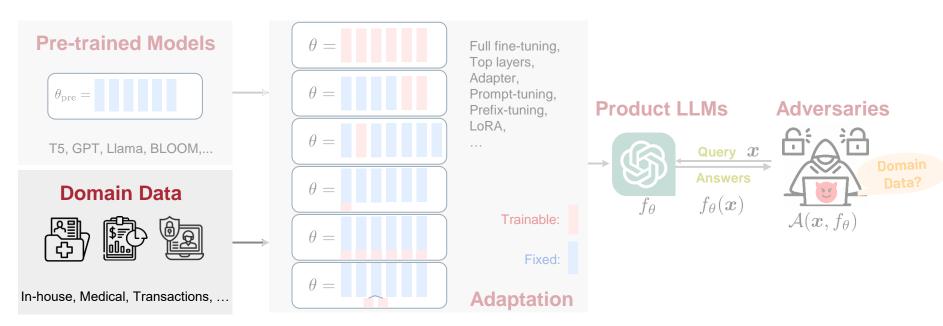


1. Adaptation techniques with diverse characteristics



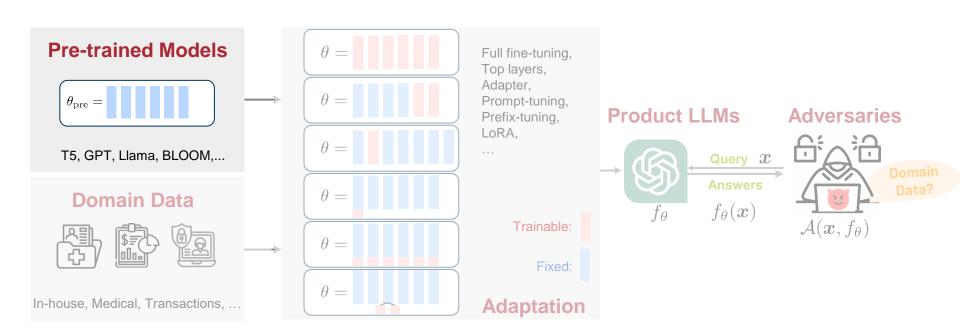


Varied domain data modalities: Finance, Corporate Climate Policy Engagement,
 Synthetic Text-to-SQL





Different pre-trained model architectures: T5, LLaMA, OPT, BLOOM, GPT-J





- 4. Representative **Attack methods**
 - Across different <u>threat models</u>:
 - ➤ White-box: Attacker has access to model internals
 - ➤ Black-box: Attacker only has access to model output probabilities (e.g., via API)



- 4. Representative Attack methods
 - Under unified notations

$$\mathcal{A}(\boldsymbol{x}, f_{\theta}) = \mathbb{1}\left[\frac{1}{L} \sum_{l=1}^{L} \log f_{\theta}(x_{l} | x_{1}, ..., x_{l-1}) > \tau_{L}\right]$$

$$\mathcal{A}(x, f_{\theta}) = \mathbb{1}\left[\frac{1}{L} \sum_{l=1}^{L} \left(\log f_{\theta}(x_{l}|x_{1}, ..., x_{l-1}) - \log f_{\phi}(x_{l}|x_{1}, ..., x_{l-1})\right) > \tau_{L_{\text{ref}}}\right]$$

$$\mathcal{A}(\boldsymbol{x}, f_{\theta}) = \mathbb{1}\left[-\frac{1}{L}\sum_{l=1}^{L}\log f_{\theta}(x_{l}|x_{1}, ..., x_{l-1})/\mathcal{H}(\boldsymbol{x}) < \tau_{\text{zlip}}\right]$$

$$\mathcal{A}(\boldsymbol{x}, f_{\theta}) = \mathbb{1} \left[\frac{1}{L} \sum_{l=1}^{L} \log f_{\theta}(x_{l} | x_{1}, ..., x_{l-1}) - \frac{1}{kL} \sum_{i=1}^{k} \sum_{l=1}^{L} \log f_{\phi}(\tilde{x}_{l}^{(i)} | \tilde{x}_{1}^{(i)}, ..., \tilde{x}_{l-1}^{(i)}) > \tau_{L_{\text{nbr}}} \right]$$

$$\mathcal{A}(\boldsymbol{x}, f_{\theta}) = \mathbb{1} \left[\frac{1}{|\text{Min-K}\%(\boldsymbol{x})|} \sum_{x_{l} \in \text{Min-K}\%(\boldsymbol{x})} \log f_{\theta}(x_{l} | x_{1}, ..., x_{l-1}) > \tau_{\text{Min-K}} \right]$$

$$\mathcal{A}(\boldsymbol{x}, f_{\theta}) = \mathbb{1}\left[\frac{1}{|\text{Min-K}\%(\boldsymbol{x})|} \sum_{\boldsymbol{x} \in \text{Min-K}\%(\boldsymbol{x})} \frac{\log f_{\theta}(x_{l}|x_{1}, ..., x_{l-1}) - \mu_{< l}}{\sigma_{< l}} > \tau_{\text{Min-K++}}\right]$$

Gradient Norm-based
$$\mathcal{A}(\boldsymbol{x}, f_{\theta}) = \mathbb{1} \Big[\big\| - \frac{1}{L} \sum_{l=1}^{L} \nabla_{\theta} \log f_{\theta}(x_{l} | x_{1}, ..., x_{l-1}) \big\| < \tau_{\text{grad}} \Big]$$



- 1. Adaptation techniques with diverse characteristics
- 2. Varied domain data modalities
- 3. Different pre-trained model architectures
- 4. Representative **Attack methods**



- 1. Adaptation techniques with diverse characteristics
- 2. Varied domain data modalities
- 3. Different pre-trained model architectures
- 4. Representative Attack methods

Research Questions

RQ 1: Is private data used for adapting LLMs vulnerable to leaks?

whether

RQ 2: Do different adaptation techniques vary in their downstream privacy vulnerability? what

RQ 3: What factors potentially affect privacy vulnerability in LLM adaptation?

how & why

RQ1: Is data vulnerable to leaks?



Distributional difference generally exist between member and non-member data.

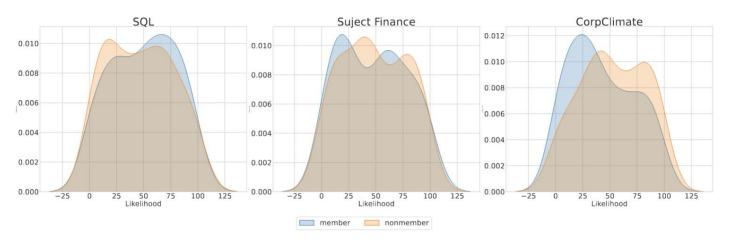


Figure 2: The likelihood score distribution of member and non-member data in Llama-7b fine-tuned with LoRA on different datasets.

RQ1: Is data vulnerable to leaks?



- **Distributional difference** generally exist between member and non-member data.
- Strong MIAs effectively detect data used for LLM adaptation.

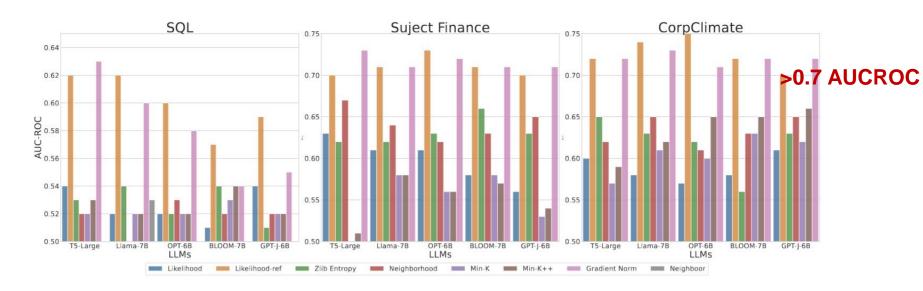
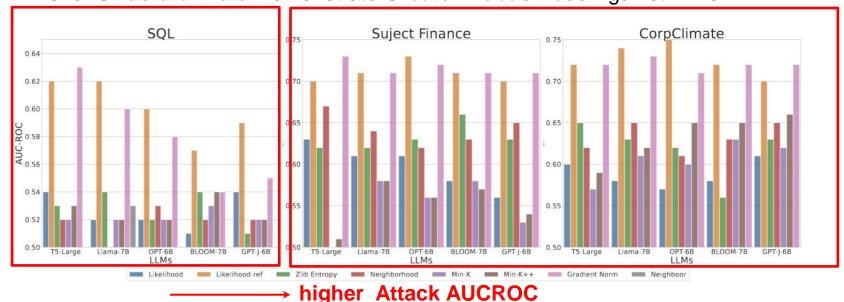


Figure 3: Overview of the attack performance across different LLMs and datasets.

RQ1: Is data vulnerable to leaks?



- Distributional difference generally exist between member and non-member data.
- Strong MIAs effectively detect data used for LLM adaptation.
- LLMs for Structural Data Demonstrate Greater Robustness Against MIAs.



RQ 2: Impact of Adaptation Techniques



More trainable parameters lead to higher data membership leakage risk.

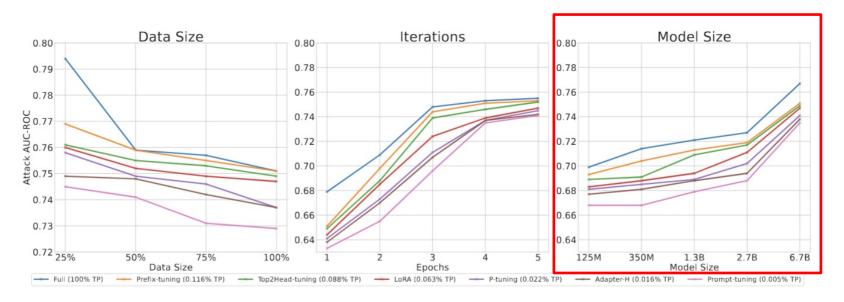


Figure 5: Impact of different adaptation techniques for *attack performance* measured by AUC-ROC. TP refers to the percentage of trainable parameters compared to the full-size model parameters.

RQ 2: Impact of Adaptation Techniques



 Different adaptation techniques may cause systematic vulnerability differences due to their associated attack surfaces.

Adaptation Method	Attack Method							Agourgou (ofter)
	Likelihood	Likelihood-ref	Zlib Entropy	Neighborhood	Min-K	Min-K++	Gradient-Norm	Accuracy (after)
Prompt-tuning	0.562	0.629	0.591	0.619	0.554	0.579	0.635	0.664
P-tuning	0.587	0.636	0.628	0.633	0.583	0.595	0.644	0.676
Prefix-tuning	0.574	0.648	0.633	0.635	0.577	0.601	0.642	0.671
Adapter-H	0.556	0.675	0.607	0.628	0.566	0.579	0.659	0.669
LoRA	0.575	0.735	0.634	0.654	0.608	0.622	0.728	0.674
Top2-head	0.677	0.788	0.714	0.694	0.647	0.696	0.793	0.669
Full	0.832	0.882	0.847	0.803	0.787	0.827	0.879	0.677
In-Context	0.922	0.922	0.922	0.922	0.922	0.922	0.922	0.534
From scratch	0.913	0.943	0.914	0.899	0.892	0.921	0.958	0.278



Utilizing more data make the attack less effective.

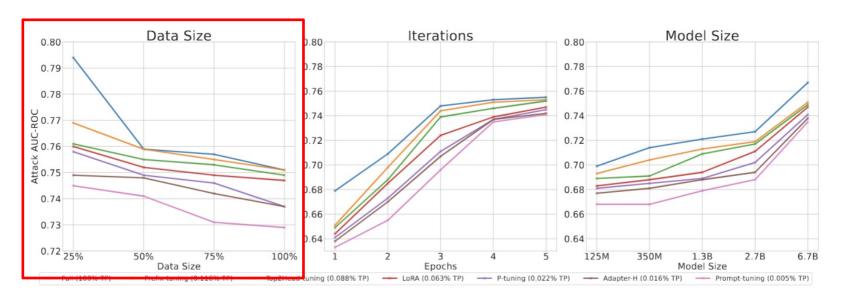


Figure 5: Impact of different adaptation techniques for *attack performance* measured by AUC-ROC. TP refers to the percentage of trainable parameters compared to the full-size model parameters.



 Increasing the <u>number of iterations</u> generally enhances the effectiveness of attacks on the target models.

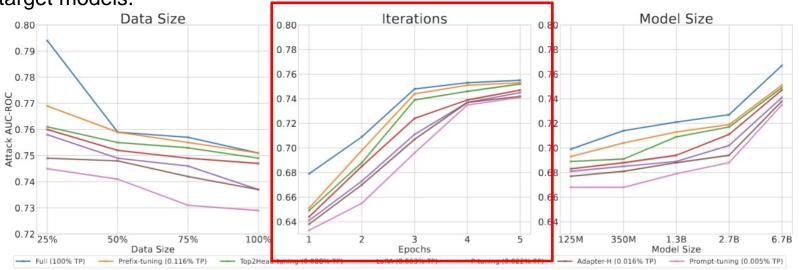


Figure 5: Impact of different adaptation techniques for *attack performance* measured by AUC-ROC. TP refers to the percentage of trainable parameters compared to the full-size model parameters.



Larger LLMs tend to exhibit increased downstream privacy vulnerability after adaptation.

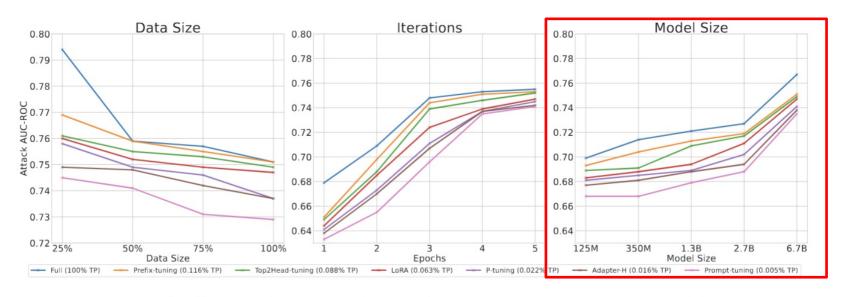


Figure 5: Impact of different adaptation techniques for *attack performance* measured by AUC-ROC. TP refers to the percentage of trainable parameters compared to the full-size model parameters.



Larger LLMs tend to exhibit increased downstream privacy vulnerability after adaptation.

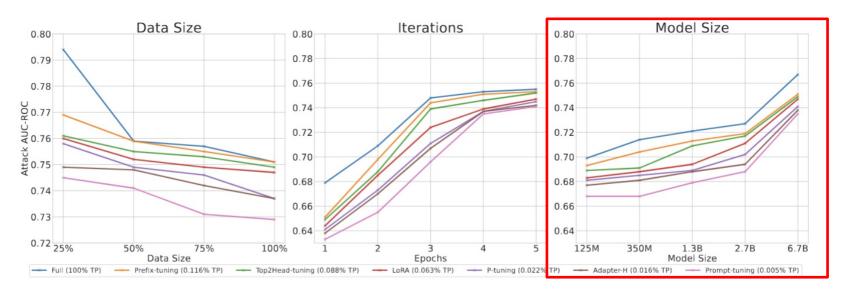


Figure 5: Impact of different adaptation techniques for *attack performance* measured by AUC-ROC. TP refers to the percentage of trainable parameters compared to the full-size model parameters.



Larger LLMs tend to exhibit increased downstream <u>privacy vulnerability</u> after adaptation.

increased downstream model utility after adaptation. Trade-of

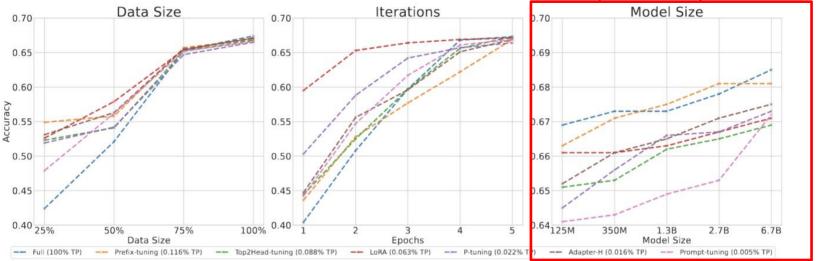


Figure 6: Impact of different adaptation techniques for *model utility* measured by accuracy. TP refers to the percentage of trainable parameters compared to the full-size model parameters.

Thank you



Github link: https://github.com/yKvD89Sri8/llm_finetuning_privacy_benchmark)