SOFT: Selective Data Obfuscation for Protecting LLM Fine-tuning against Membership Inference Attacks

Kaiyuan Zhang, Siyuan Cheng, Hanxi Guo, Yuetian Chen, Zian Su, Shengwei An, Yuntao Du, Charles Fleming†, Ashish Kundu†, Xiangyu Zhang, Ninghui Li

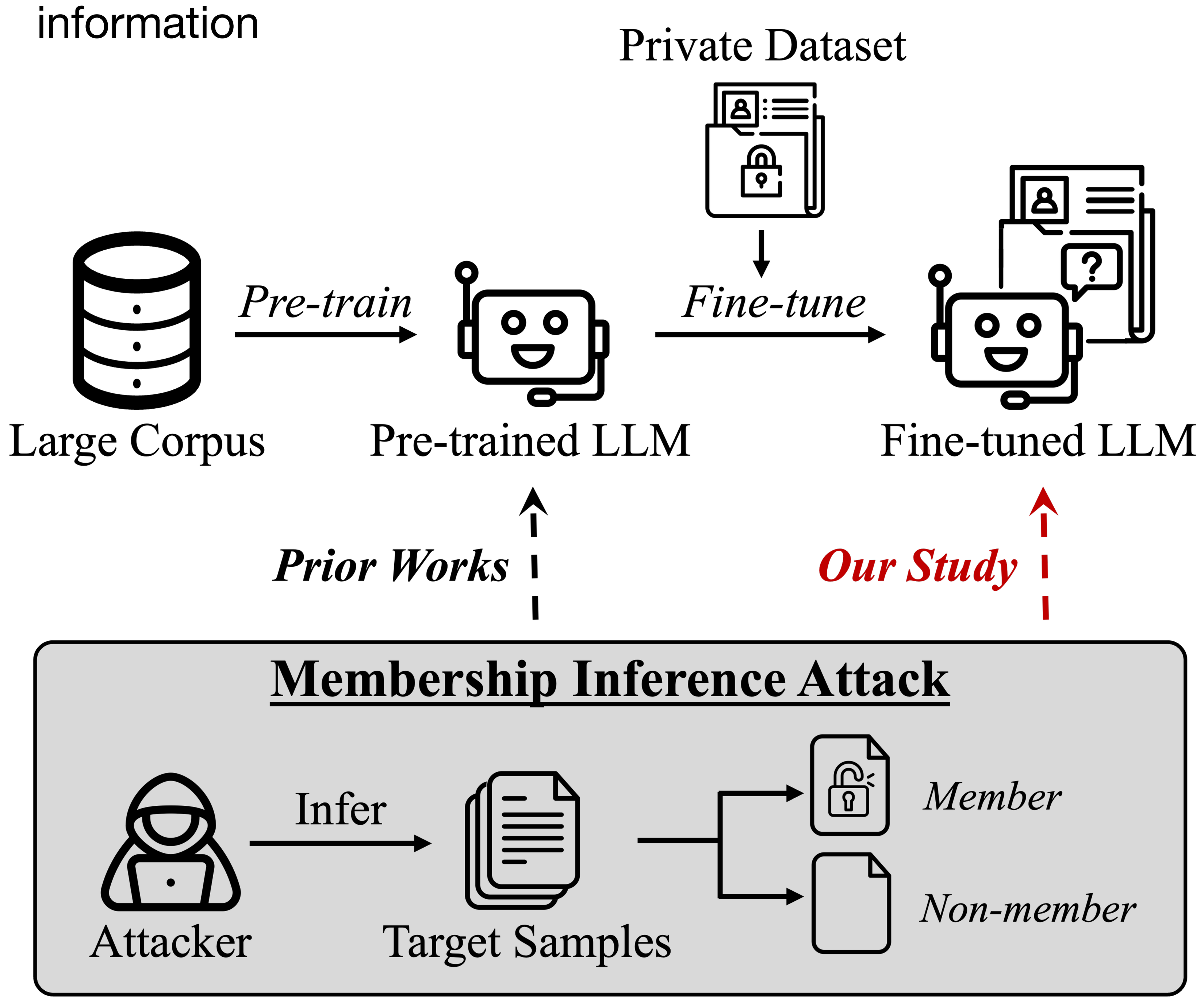
PURDUE UNIVERSITY

† CISCO

34TH USENIX SECURITY SYMPOSIUM

Problem & Motivation

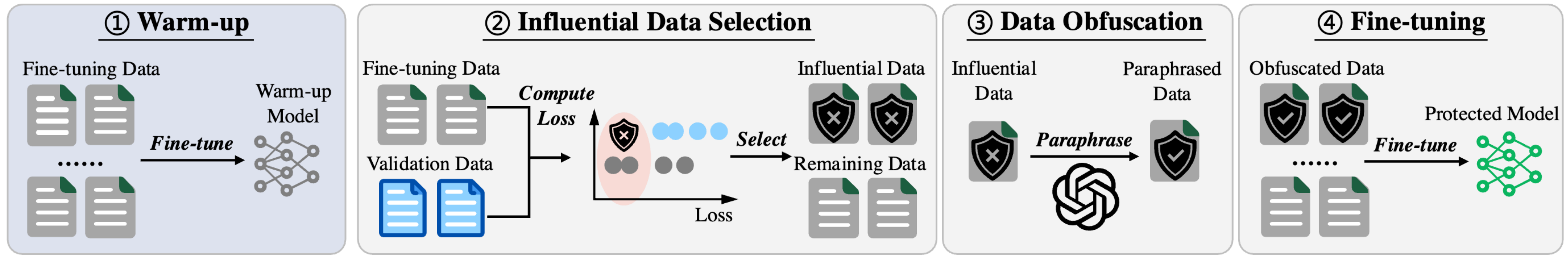
- MIA determines *whether a specific data record was used* to train a target model or not
- Pre-training large-scale LLMs requires resources, e.g. A100 GPUs
- Small companies and individuals use pre-trained model as the backbone to fine-tune
- Data used in **fine-tuning** often includes either PII, copyright data, or even confidential organizational information



The Calibration Challenge

The Calibration Challenge. *Existing LLM MIAs mainly differ on how to differentiate **uncommon** sentences used in training from **common** sentences not used in training. Many of these methods share similarities on **calibration** and differ mainly in their use of loss, log-likelihood, perplexity, contrastive ratios, or an extra reference model.*

Selective Data Obfuscation Overview



Observation

	ArXiv	DM Math.	HackerNews	PubMed	Pile CC	Wikipedia	GitHub
Loss	0.878	0.647	0.854	0.848	0.863	0.853	0.889
Zlib	0.882	0.609	0.861	0.850	0.859	0.862	0.909
Lowercase	0.843	0.589	0.861	0.811	0.833	0.830	0.885
Min-K% Prob	0.650	0.542	0.631	0.624	0.674	0.617	0.604
Min-K%++	0.866	0.618	0.827	0.842	0.850	0.841	0.908
Ratio	0.874	0.773	0.866	0.865	0.863	0.875	0.922
Bag of words	0.583	0.517	0.519	0.560	0.496	0.524	0.706
ReCall	0.884	0.649	0.873	0.860	0.864	0.847	0.900
CON-ReCall	0.825	0.600	0.835	0.851	0.822	0.831	0.882
Ensemble	0.872	0.710	0.873	0.871	0.868	0.876	0.839

Figure I: AUC-ROC on Full Fine-tuned Pythia

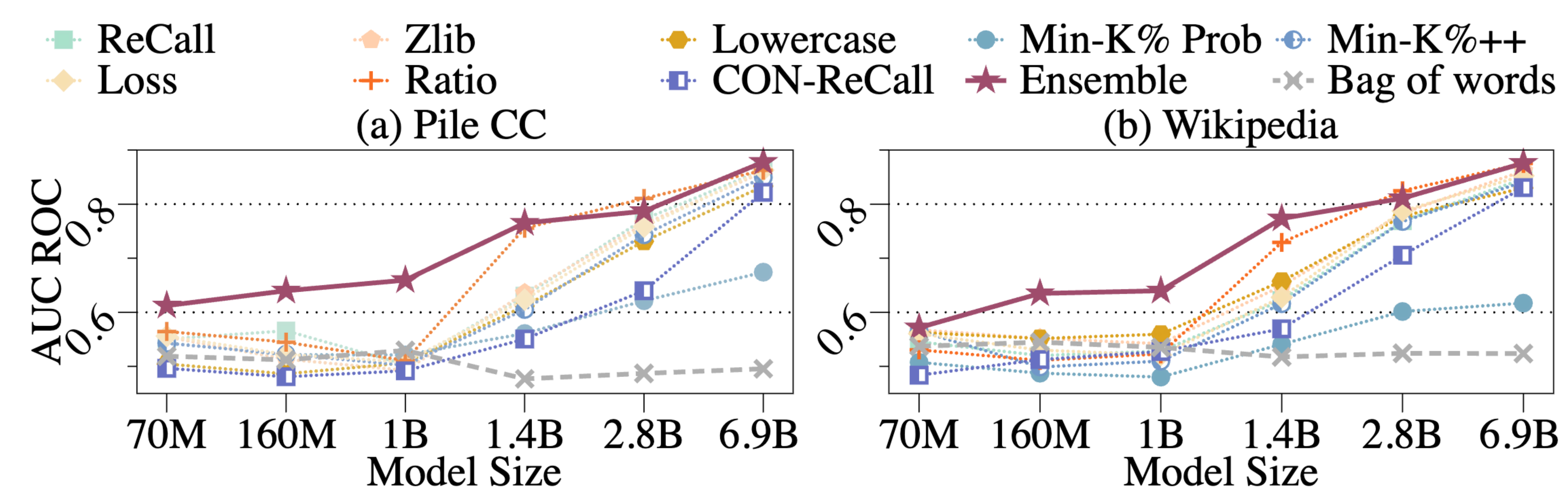


Figure II: Full Fine-tune on Different Model Sizes of Pythia

	ArXiv	DM Math.	HackerNews	PubMed	Pile CC	Wikipedia	GitHub
Loss	0.601	0.533	0.560	0.557	0.527	0.571	0.770
Zlib	0.599	0.524	0.569	0.548	0.514	0.583	0.766
Lowercase	0.578	0.501	0.561	0.538	0.533	0.595	0.772
Min-K% Prob	0.602	0.547	0.544	0.519	0.544	0.527	0.752
Min-K%++	0.591	0.523	0.546	0.562	0.544	0.544	0.762
Ratio	0.628	0.549	0.634	0.613	0.590	0.644	0.803
Bag of words	0.597	0.469	0.504	0.529	0.529	0.527	0.700
ReCall	0.611	0.523	0.575	0.547	0.532	0.577	0.755
CON-ReCall	0.592	0.530	0.544	0.466	0.546	0.562	0.768
Ensemble	0.663	0.625	0.623	0.666	0.618	0.637	0.807

Figure III: AUC-ROC on LoRA Fine-tuned Pythia

Evaluation

Table I: Evaluation of SOFT in AUC-ROC Score

MIAs	ArXiv			
	Pretrain	FT	LoRA	SOFT
Loss [92]	0.508	0.822	0.601	0.525
Zlib [16]	0.508	0.811	0.593	0.521
Lowercase [16]	0.490	0.785	0.577	0.517
Min-K% Prob [73]	0.514	0.615	0.554	0.510
Min-K%++ [98]	0.509	0.757	0.584	0.519
Ratio [16]	0.493	0.952	0.689	0.558
Bag of words [62]	0.504	0.508	0.508	0.505
ReCall [87]	0.508	0.840	0.582	0.533
CON-ReCall [82]	0.505	0.764	0.557	0.518
Ensemble	0.551	0.807	0.663	0.568
Average	0.509	0.766	0.591	0.527

Table II: Adaptive Attacks

Setting	AUC-ROC	TPR@1%FPR
No Defense (FT)	0.807	0.258
Paraphrase & Selection	0.595	0.149
Paraphrase Only	0.575	0.136
Selection Only	0.651	0.086
No Adaptive (w/ SOFT)	0.568	0.033

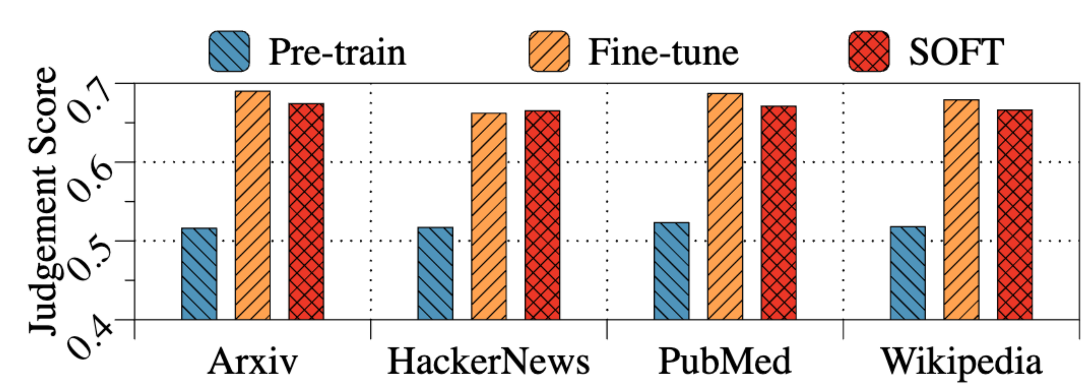


Figure IV: Utility test using LLM-as-a-Judge

Project Page

Personal Page