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

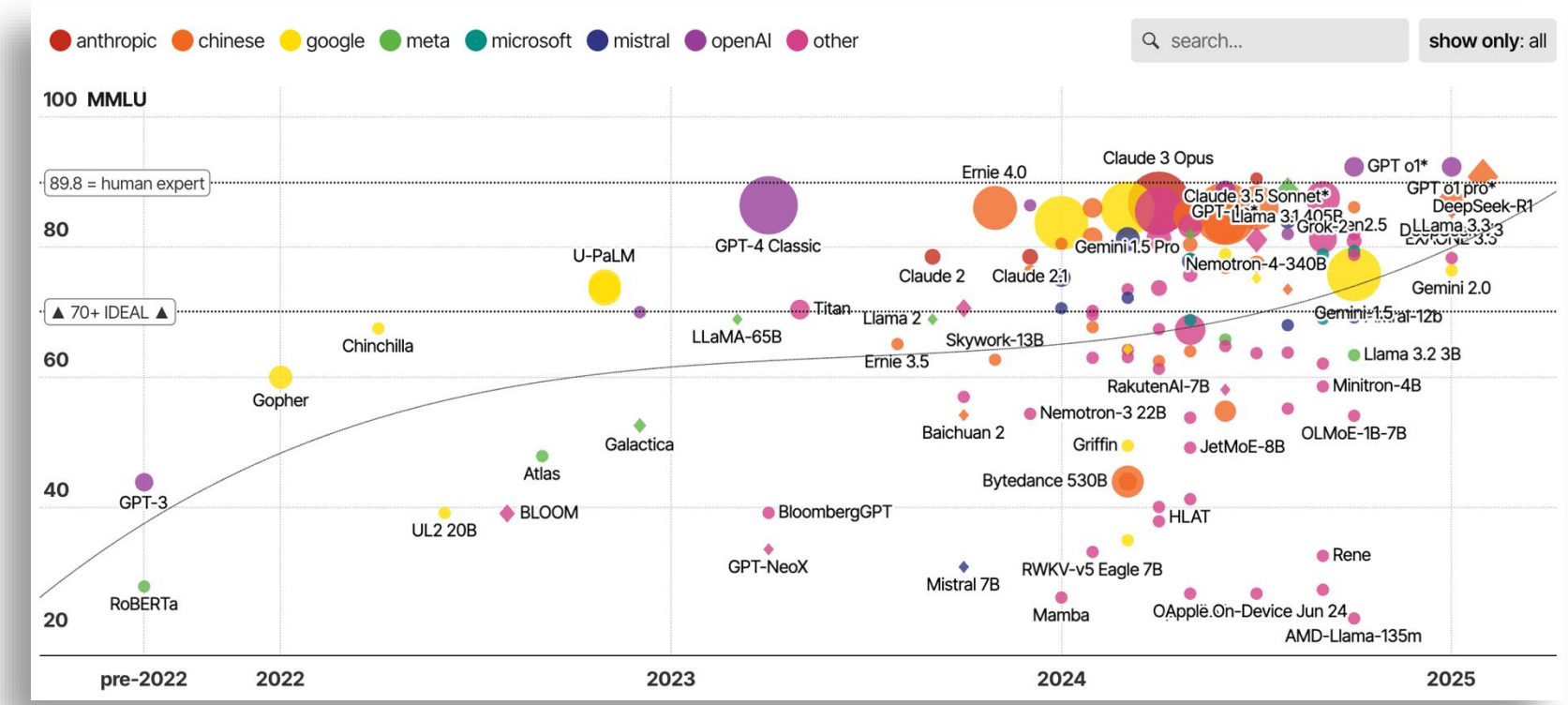
Kaiyuan Zhang, Siyuan Cheng, Hanxi Guo, Yuetian Chen, Zian Su, Shengwei An, Yuntao Du, Charles Fleming†, Ashish Kundu†, Xiangyu Zhang, Ninghui Li

USENIX Security 2025

LLM Advancement

Major Large Language Models (LLMs)

ranked by capabilities, sized by billion parameters used for training

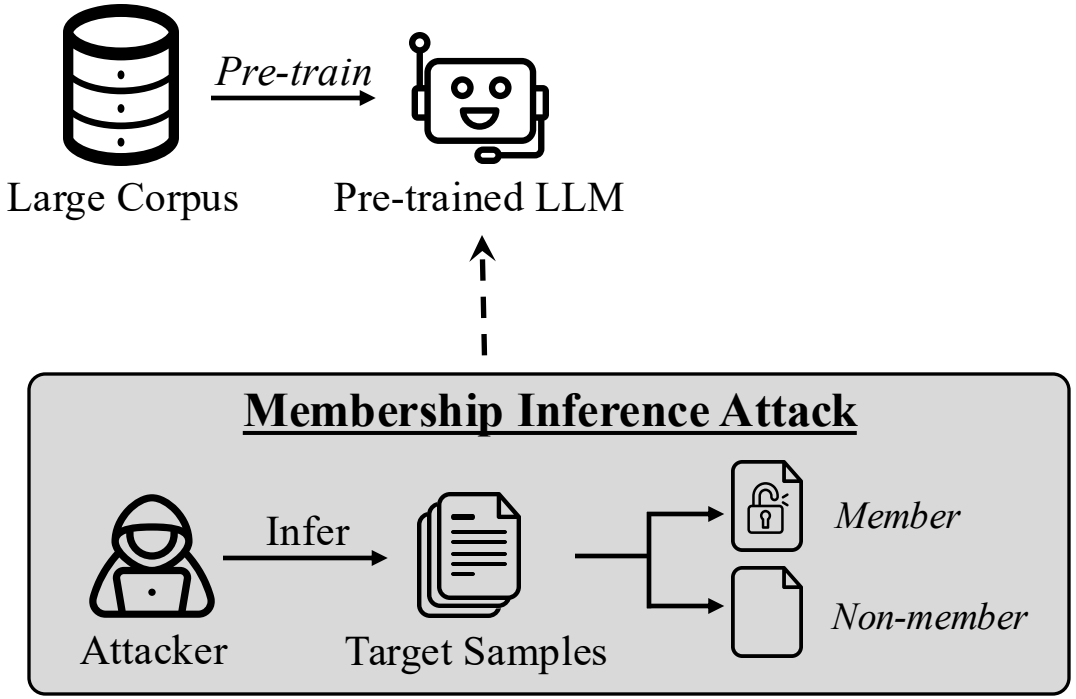


[1]. Source: informationisbeautiful.net

Membership Inference Attack

MIA determines *whether a specific data record was used* to train a target model or not

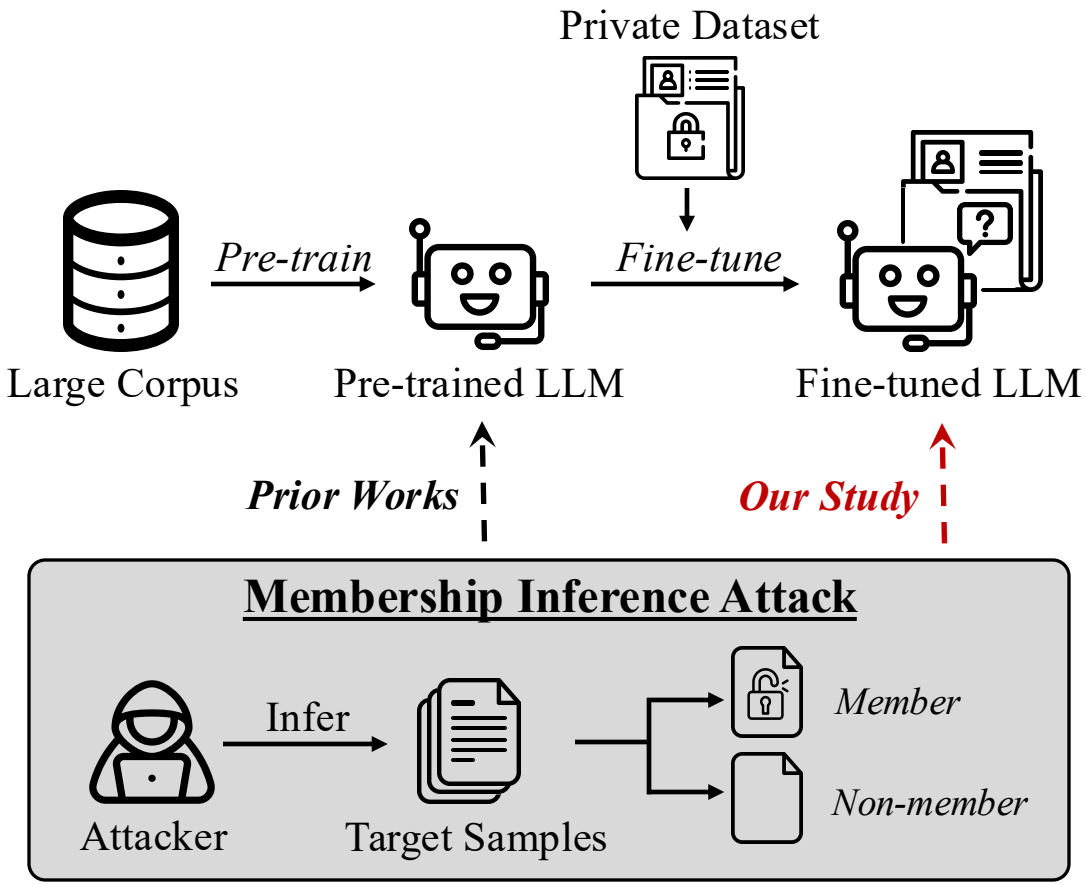
- LLM pre-training
  - Pre-training large-scale LLMs requires resources, e.g. A100 GPUs
  - Small companies and individuals use pre-trained model as the backbone to fine-tune



Membership Inference Attack

MIA determines *whether a specific data record was used* to train a target model or not

- LLM pre-training
- LLM fine-tuning
  - 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.*

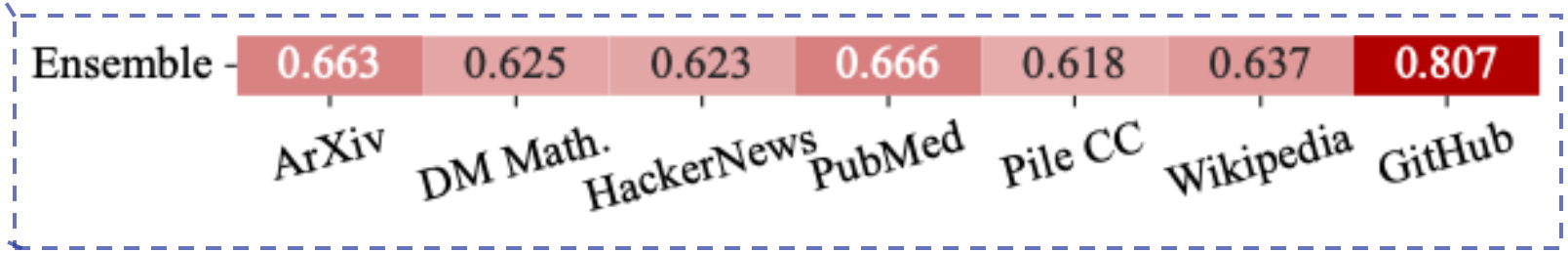
- The ineffectiveness of existing membership inference attacks in pre-trained LLMs, motivating the introduce of the ***Ensemble attack***.

Pitfalls in Fine-tuning

- As model size and fine-tune epoch increase, *fully fine-tuned* LLMs exhibit *greater privacy leakage*.
- Even *one-epoch* fine-tuning results in significant leakage.

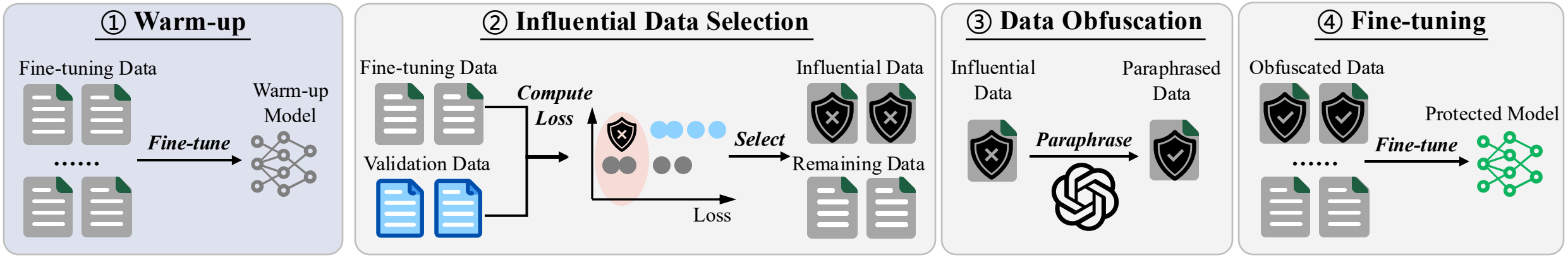
Privacy-Utility Trade-offs in LoRA

- Comparing LoRA with full fine-tuning, while LoRA achieves a better trade-off between privacy and utility, the *ensemble* and *ratio attack* remain capable of compromising it.



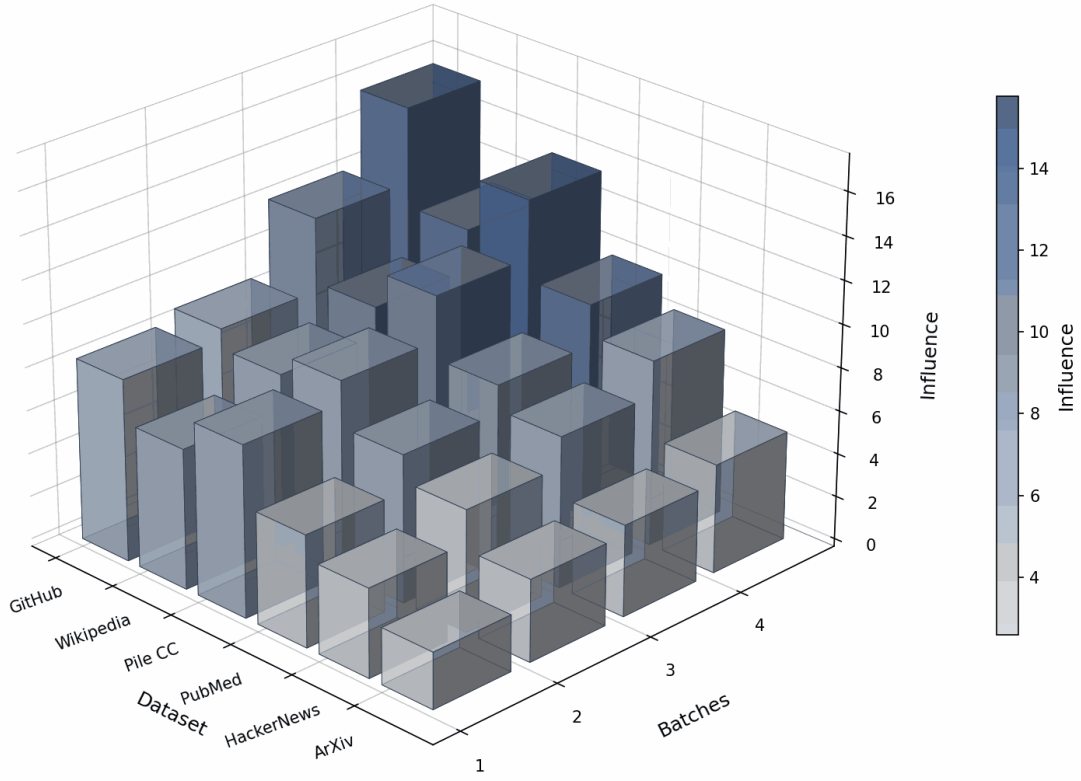
Selective Data Obfuscation

- In high-level, SOFT involves *substituting selective influential samples* with *semantically equivalent alternatives* by a paraphraser during fine-tuning.



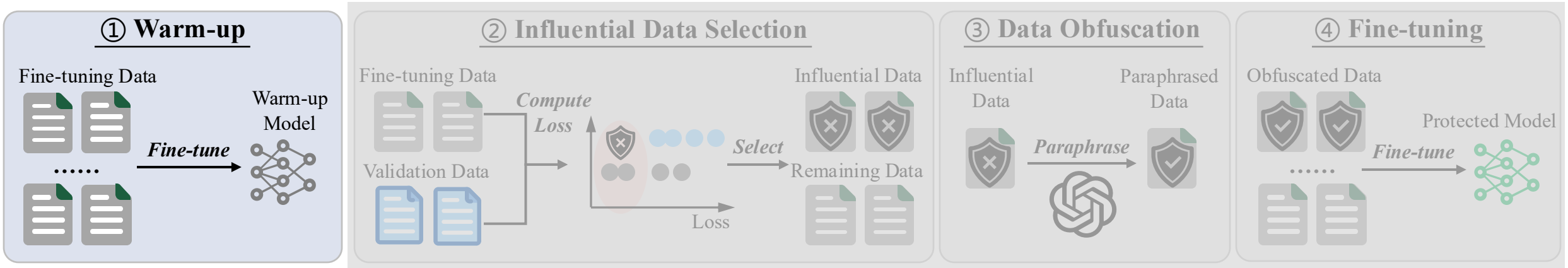
Data Selection

- Inspired by influence function [2], we define influential samples as those vulnerable to MIA.
- SOFT selectively replaces influential samples, i.e., *those are easily memorized and exhibit lower loss values*, with their obfuscated counterparts.



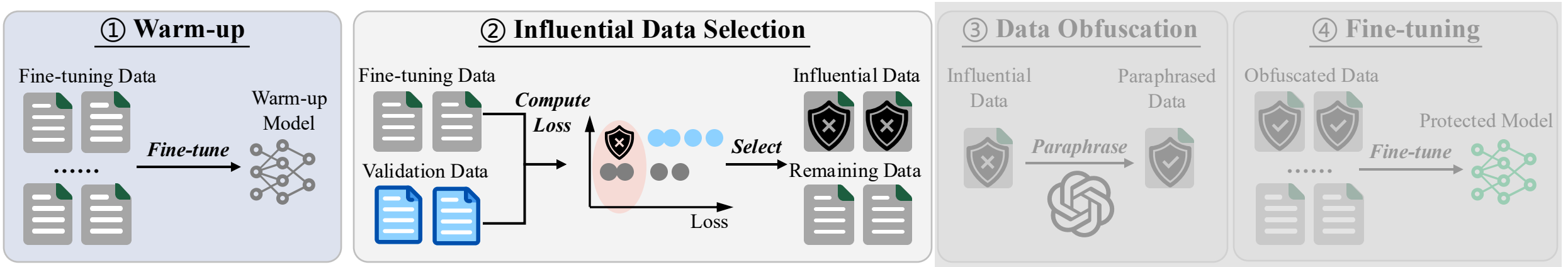
[2]. Koh, Pang Wei, and Percy Liang. Understanding black-box predictions via influence functions. ICML 2017

Selective Data Obfuscation



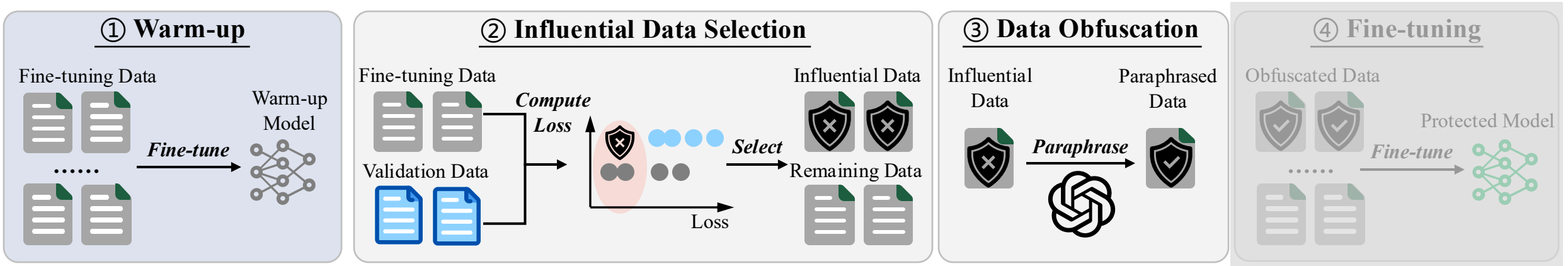
1. Warm-up Fine-tuning
   - Warm-up helps assess the initial influence level of each sample
2. Influential Data Selection
   - SOFT evaluates sample from the fine-tuning dataset and select influential ones

Selective Data Obfuscation



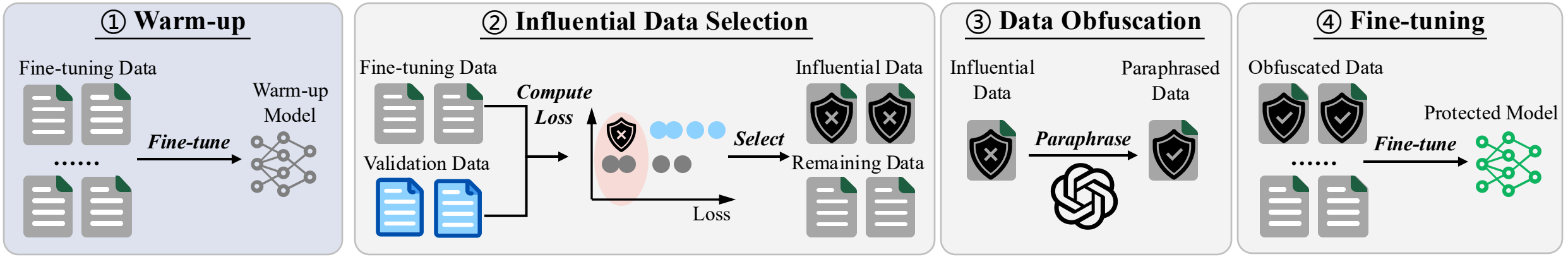
1. Warm-up Fine-tuning
   - Warm-up helps assess the initial influence level of each sample
2. Influential Data Selection
   - SOFT evaluates sample from the fine-tuning dataset and select influential ones

Selective Data Obfuscation



3. Data Obfuscation
   - SOFT replaces the selected influential samples with paraphrased versions
4. Fine-tuning
   - Combining the obfuscated data with the remaining safe data, SOFT fine-tunes on the updated dataset

Selective Data Obfuscation



3. Data Obfuscation
   - SOFT replaces the selected influential samples with paraphrased versions
4. Fine-tuning
   - Combining the obfuscated data with the remaining safe data, SOFT fine-tunes on the updated dataset

Evaluation

- Does SOFT effective in defending against MIAs?

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 1: **Evaluation of SOFT's defense effectiveness against multiple MIAs, with comparison to LoRA and full fine-tuning (FT).** Performance is measured using **AUC-ROC** scores, where lower values indicate stronger defense.

MIAs	ArXiv				HackerNews				PubMed				Pile CC				Wikipedia				GitHub			
	Pretrain	FT	LoRA	SOFT	Pretrain	FT	LoRA	SOFT	Pretrain	FT	LoRA	SOFT	Pretrain	FT	LoRA	SOFT	Pretrain	FT	LoRA	SOFT	Pretrain	FT	LoRA	SOFT
Loss [92]	0.508	0.822	0.601	0.525	0.498	0.900	0.645	0.515	0.478	0.895	0.619	0.496	0.502	0.887	0.633	0.519	0.501	0.936	0.644	0.530	0.653	0.846	0.750	0.625
Zlib [16]	0.508	0.811	0.593	0.521	0.496	0.910	0.641	0.517	0.481	0.893	0.621	0.509	0.489	0.902	0.648	0.533	0.505	0.939	0.644	0.532	0.678	0.871	0.776	0.647
Lowercase [16]	0.490	0.785	0.577	0.517	0.507	0.845	0.575	0.515	0.515	0.850	0.595	0.541	0.482	0.858	0.598	0.522	0.499	0.887	0.650	0.536	0.611	0.820	0.716	0.591
Min-K% Prob [73]	0.514	0.615	0.554	0.510	0.492	0.627	0.541	0.489	0.502	0.645	0.550	0.499	0.511	0.668	0.547	0.518	0.495	0.669	0.638	0.512	0.506	0.613	0.643	0.515
Min-K%++ [98]	0.509	0.757	0.584	0.519	0.498	0.800	0.579	0.511	0.486	0.856	0.568	0.503	0.507	0.842	0.549	0.518	0.519	0.912	0.744	0.533	0.606	0.869	0.640	0.598
Ratio [16]	0.493	0.952	0.689	0.558	0.462	0.943	0.702	0.533	0.503	0.947	0.692	0.541	0.510	0.949	0.918	0.552	0.488	0.944	0.774	0.576	0.507	0.955	0.922	0.516
Bag of words [62]	0.504	0.508	0.508	0.505	0.529	0.521	0.521	0.523	0.513	0.528	0.528	0.518	0.483	0.504	0.511	0.511	0.501	0.507	0.507	0.507	0.701	0.649	0.651	0.660
ReCall [87]	0.508	0.840	0.582	0.533	0.501	0.907	0.542	0.515	0.480	0.908	0.547	0.511	0.497	0.895	0.545	0.532	0.505	0.938	0.641	0.529	0.630	0.851	0.750	0.627
CON-ReCall [82]	0.505	0.764	0.557	0.518	0.486	0.740	0.577	0.500	0.488	0.868	0.556	0.516	0.458	0.844	0.557	0.513	0.496	0.925	0.627	0.530	0.638	0.847	0.743	0.620
Ensemble	0.551	0.807	0.663	0.568	0.524	0.886	0.749	0.567	0.576	0.884	0.653	0.546	0.673	0.942	0.884	0.604	0.512	0.925	0.847	0.587	0.747	0.944	0.858	0.669
Average	0.509	0.766	0.591	0.527	0.499	0.808	0.607	0.519	0.502	0.827	0.593	0.518	0.511	0.829	0.639	0.532	0.502	0.858	0.672	0.537	0.628	0.827	0.745	0.607

Observations: SOFT effectively reduces attack efficacy by significantly lowering the AUC-ROC scores to 0.527 on ArXiv.

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

Take-aways:

1. SOFT is designed to protect LLM fine-tuning against membership inference attacks.
2. SOFT is grounded in influence functions and data selection.
3. SOFT selectively replaces influential samples with their obfuscated counterparts.
4. Paper, code, slides: https://soft-mia.github.io/

Thank you for listening!

SCAN ME

PURDUE UNIVERSITY CISCO