Membership Inference Attacks Against Fine-tuned Diffusion Language Models

Yuetian Chen[1], Kaiyuan Zhang[1], Yuntao Du[1], Edoardo Stoppa[1], Charles Fleming[2], Ashish Kundu[2], Bruno Ribeiro[1], Ninghui Li[1]

[1] Department of Computer Science, Purdue University;

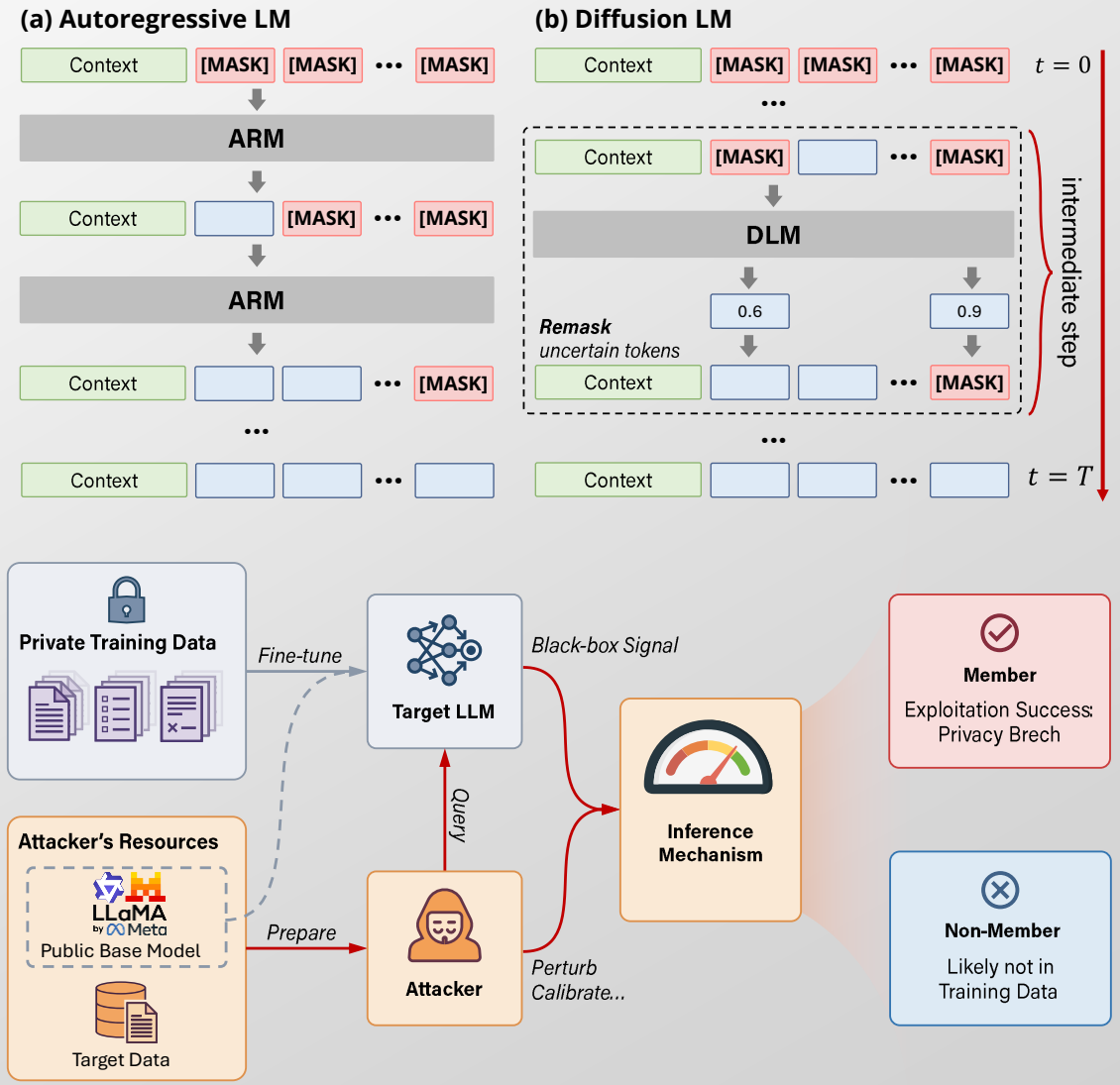
[2] Cisco Research

Presenter: Yuetian Chen

Context: Privacy Risks in Diffusion Language Models

Do the unique paradigms of DLMs introduce novel vulnerabilities exploitable by MIAs?

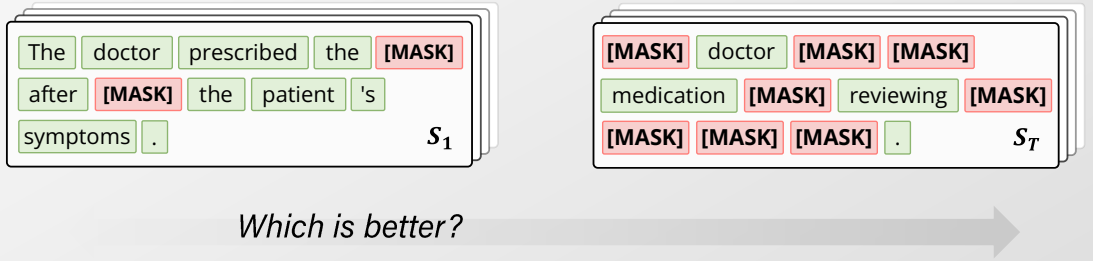
- **A Distinct Paradigm: Diffusion Language Models (DLMs)**
  - DLMs are emerging as a compelling alternative to Autoregressive Models (ARMs)
  - **Key Difference:** Instead of fixed sequential prediction, DLMs use bidirectional context to iteratively reconstruct masked tokens
- **The Threat:** Membership Inference Attacks (MIA)
  - **Goal:** Detect if a specific text sample (e.g., private user data) was used to fine-tune the target model
  - **Threat Model:** We assume a standard Grey-box adversary with query access (logits) to the fine-tuned model and access to a reference model
- **The Opportunity: Exploiting "Extra" Signals**
  - The masking mechanism allows attackers to probe multiple, arbitrary configurations of the same text in DLMs
  - **Hypothesis:** This flexibility exposes richer, multi-view membership signals that existing attacks fail to capture



The Challenges of Attacking DLMs

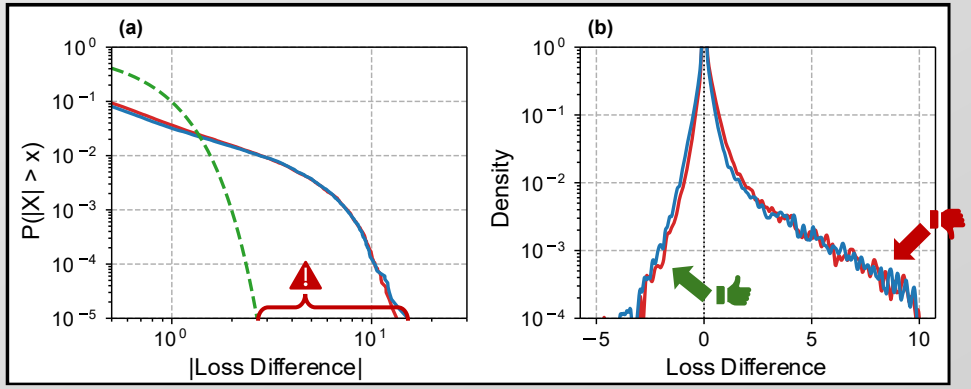
How we mask? & How we score?

- **The Masking Dilemma**
  - A random mask configuration will likely miss the specific token pattern the model memorized
  - You mostly get weak signals dominated by pure noise



- **The Problem with Loss Averaging**
  - Long tail noise make loss averaging unreliable
  - **Domain tokens are "loud":** Extreme token loss spikes usually come from ubiquitous "domain adaptation" terms (e.g., “Biography")
  - **Contamination:** These loss outliers contain **little membership info**, yet their sheer magnitude completely dominates the average score

Most Frequent Token	Loss Reduction after Fine-tune	Membership Signal Strength	Count in Corpus
‘ Biography’	1.040	0.295	50
‘ Background’	0.671	0.027	25
‘ Life’	0.631	0.299	21
‘ List’	0.544	0.499	39
‘ Description’	0.399	0.474	23
‘ Joseph’	0.393	0.022	23



Methodology: SAMA

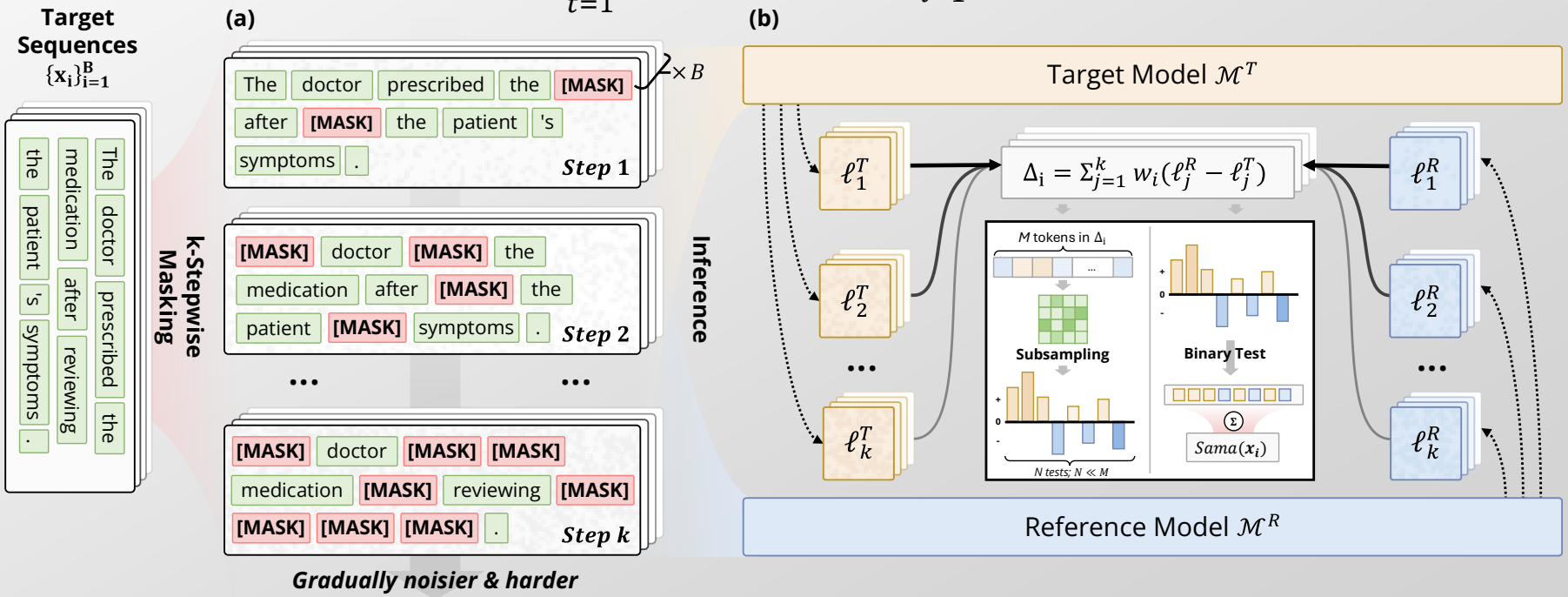
Subset-Aggregated Membership Attack

- **(a) Solving The Masking Dilemma - Progressive Evidence Collection**
  - **Memorization is a moving target.** We triangulate memorization across multiple densities.
  - **Multi-Scale Probing:** Systematically sweep masking density α_t over T steps to capture patterns at every context level:

$$\alpha_t = \alpha_{min} + \frac{t-1}{T-1}(\alpha_{max} - \alpha_{min})$$

  - **Adaptive Weighting:** Inverse-step weighting automatically prioritizes the cleaner, high-fidelity signals found in sparse masks:

$$SAMA(x) = \sum_{t=1}^{T} w_t \hat{\beta}_t(x), \quad w_t = \frac{1/t}{\sum_{i=1}^{T} 1/i}$$



Methodology: SAMA

Subset-Aggregated Membership Attack

- **(b) Solving Long-tail Noise - Robust Subset Aggregation**
  - Sign-based estimators neutralize infinite-variance domain noise, isolating sparse memorization signals.
  - **The Statistical Dilemma:** Standard Monte Carlo estimates of the expected loss gap, $\mathbb{E}_{\mathbb{S}}[\Delta_{DF}(x;\mathcal{S})]$, are easily hijacked by heavy-tailed domain adaptation artifacts (excess kurtosis > 80).

  - **Localized Subsampling:** We isolate extreme outliers by drawing N random subsets $\mathcal{U}^n \subset \mathcal{S}_t$ to compute localized loss gaps:

$$\Delta_{DF}^{n}(x;\mathcal{S}_t) = \frac{1}{m}\sum_{i\in\mathcal{U}^n}\left[l_i^R(x,\mathcal{S}_t) - l_i^T(x,\mathcal{S}_t)\right]$$

  - **Sign-Based Transformation:** Leveraging Hodges-Lehmann robust statistics, we map the continuous gap to a binary indicator, discarding magnitude entirely

$$B^n(x) = \mathbf{1}[\Delta_{DF}^{n}(x;\mathcal{S}_t) > 0]$$

  - **The Theoretical Guarantee:** For non-members, symmetric noise ensures $\mathbb{E}[B^n] = 0.5$ regardless of infinite variance. For members, consistent memorization breaks this symmetry, dominating the aggregated vote $\widehat{\beta}_t(x) = \frac{1}{N}\sum B^n(x)$.

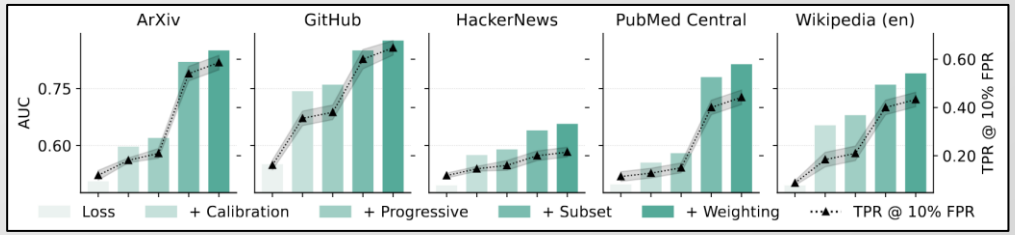
Experimental Results

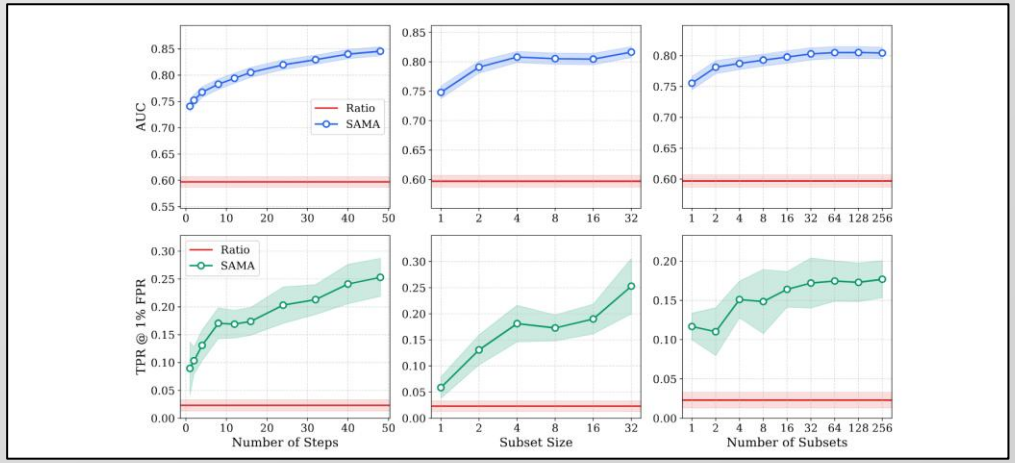
Table 1: MIA performance (AUC, TPR@10%FPR, TPR@1%FPR, TPR@0.1%FPR) across datasets. Each cell shows the mean with std. dev. as a subscript. The best results are highlighted.

MIAs	ArXiv				GitHub				HackerNews			
	AUC	T@10%	T@1%	T@0.1%	AUC	T@10%	T@1%	T@0.1%	AUC	T@10%	T@1%	T@0.1%
Loss	$0.506_{\pm.01}$	$0.119_{\pm.02}$	$0.010_{\pm.01}$	$0.000_{\pm.00}$	$0.551_{\pm.01}$	$0.161_{\pm.02}$	$0.036_{\pm.01}$	$0.007_{\pm.01}$	$0.495_{\pm.01}$	$0.118_{\pm.01}$	$0.010_{\pm.01}$	$0.000_{\pm.00}$
ZLIB	$0.490_{\pm.01}$	$0.119_{\pm.02}$	$0.012_{\pm.00}$	$0.000_{\pm.00}$	$0.561_{\pm.01}$	$0.205_{\pm.02}$	$0.045_{\pm.01}$	$0.007_{\pm.00}$	$0.486_{\pm.01}$	$0.083_{\pm.01}$	$0.009_{\pm.01}$	$0.001_{\pm.00}$
Lowercase	$0.515_{\pm.01}$	$0.103_{\pm.01}$	$0.013_{\pm.00}$	$0.001_{\pm.00}$	$0.579_{\pm.01}$	$0.178_{\pm.02}$	$0.044_{\pm.01}$	$0.008_{\pm.00}$	$0.483_{\pm.01}$	$0.074_{\pm.01}$	$0.007_{\pm.00}$	$0.001_{\pm.00}$
Min-K%	$0.488_{\pm.01}$	$0.102_{\pm.01}$	$0.012_{\pm.00}$	$0.001_{\pm.00}$	$0.530_{\pm.01}$	$0.171_{\pm.02}$	$0.039_{\pm.01}$	$0.007_{\pm.01}$	$0.492_{\pm.01}$	$0.108_{\pm.01}$	$0.013_{\pm.01}$	$0.001_{\pm.00}$
Min-K%++	$0.485_{\pm.01}$	$0.095_{\pm.01}$	$0.006_{\pm.00}$	$0.000_{\pm.00}$	$0.496_{\pm.01}$	$0.117_{\pm.01}$	$0.016_{\pm.01}$	$0.003_{\pm.00}$	$0.486_{\pm.01}$	$0.100_{\pm.02}$	$0.008_{\pm.00}$	$0.002_{\pm.00}$
BoWs	$0.519_{\pm.01}$	$0.107_{\pm.01}$	$0.011_{\pm.00}$	$0.000_{\pm.00}$	$0.656_{\pm.02}$	$0.306_{\pm.02}$	$0.154_{\pm.02}$	$0.059_{\pm.05}$	$0.527_{\pm.01}$	$0.128_{\pm.01}$	$0.009_{\pm.00}$	$0.001_{\pm.00}$
ReCall	$0.501_{\pm.01}$	$0.132_{\pm.02}$	$0.007_{\pm.00}$	$0.000_{\pm.00}$	$0.562_{\pm.01}$	$0.187_{\pm.01}$	$0.037_{\pm.01}$	$0.007_{\pm.00}$	$0.494_{\pm.01}$	$0.090_{\pm.01}$	$0.010_{\pm.01}$	$0.000_{\pm.00}$
CON-Recall	$0.500_{\pm.02}$	$0.101_{\pm.02}$	$0.011_{\pm.00}$	$0.000_{\pm.00}$	$0.549_{\pm.01}$	$0.168_{\pm.01}$	$0.027_{\pm.01}$	$0.005_{\pm.00}$	$0.501_{\pm.02}$	$0.098_{\pm.01}$	$0.015_{\pm.01}$	$0.000_{\pm.00}$
Neighbor	$0.506_{\pm.01}$	$0.098_{\pm.01}$	$0.012_{\pm.01}$	$0.001_{\pm.00}$	$0.478_{\pm.01}$	$0.072_{\pm.01}$	$0.008_{\pm.01}$	$0.000_{\pm.00}$	$0.520_{\pm.01}$	$0.123_{\pm.01}$	$0.009_{\pm.00}$	$0.001_{\pm.00}$
Ratio	$0.597_{\pm.01}$	$0.181_{\pm.01}$	$0.023_{\pm.01}$	$0.001_{\pm.00}$	$0.743_{\pm.01}$	$0.355_{\pm.03}$	$0.081_{\pm.02}$	$0.017_{\pm.01}$	$0.575_{\pm.02}$	$0.146_{\pm.01}$	$0.013_{\pm.01}$	$0.000_{\pm.00}$
SecMI	$0.520_{\pm.01}$	$0.096_{\pm.02}$	$0.006_{\pm.00}$	$0.001_{\pm.00}$	$0.604_{\pm.01}$	$0.190_{\pm.01}$	$0.044_{\pm.01}$	$0.019_{\pm.01}$	$0.523_{\pm.02}$	$0.125_{\pm.02}$	$0.013_{\pm.00}$	$0.001_{\pm.00}$
PIA	$0.525_{\pm.01}$	$0.099_{\pm.01}$	$0.011_{\pm.01}$	$0.000_{\pm.00}$	$0.571_{\pm.01}$	$0.131_{\pm.01}$	$0.012_{\pm.00}$	$0.001_{\pm.00}$	$0.494_{\pm.01}$	$0.086_{\pm.01}$	$0.014_{\pm.00}$	$0.000_{\pm.00}$
SAMA (Ours)	$\mathbf{0.850}_{\pm.01}$	$\mathbf{0.586}_{\pm.03}$	$\mathbf{0.178}_{\pm.03}$	$\mathbf{0.014}_{\pm.01}$	$\mathbf{0.876}_{\pm.01}$	$\mathbf{0.647}_{\pm.03}$	$\mathbf{0.259}_{\pm.05}$	$\mathbf{0.075}_{\pm.05}$	$\mathbf{0.657}_{\pm.01}$	$\mathbf{0.215}_{\pm.02}$	$\mathbf{0.027}_{\pm.01}$	$\mathbf{0.003}_{\pm.00}$

MIAs	PubMed Central				Wikipedia (en)				Pile CC			
	AUC	T@10%	T@1%	T@0.1%	AUC	T@10%	T@1%	T@0.1%	AUC	T@10%	T@1%	T@0.1%
Loss	$0.498_{\pm.01}$	$0.114_{\pm.02}$	$0.016_{\pm.01}$	$0.001_{\pm.00}$	$0.495_{\pm.01}$	$0.087_{\pm.00}$	$0.010_{\pm.00}$	$0.003_{\pm.00}$	$0.502_{\pm.01}$	$0.105_{\pm.01}$	$0.012_{\pm.00}$	$0.002_{\pm.00}$
ZLIB	$0.488_{\pm.01}$	$0.096_{\pm.02}$	$0.005_{\pm.00}$	$0.001_{\pm.00}$	$0.495_{\pm.01}$	$0.093_{\pm.01}$	$0.008_{\pm.00}$	$0.000_{\pm.00}$	$0.491_{\pm.01}$	$0.095_{\pm.01}$	$0.007_{\pm.00}$	$0.001_{\pm.00}$
Lowercase	$0.502_{\pm.01}$	$0.096_{\pm.01}$	$0.002_{\pm.00}$	$0.000_{\pm.00}$	$0.535_{\pm.01}$	$0.107_{\pm.02}$	$0.013_{\pm.00}$	$0.001_{\pm.00}$	$0.518_{\pm.01}$	$0.102_{\pm.01}$	$0.009_{\pm.00}$	$0.001_{\pm.00}$
Min-K%	$0.500_{\pm.01}$	$0.119_{\pm.01}$	$0.008_{\pm.00}$	$0.002_{\pm.00}$	$0.482_{\pm.01}$	$0.070_{\pm.01}$	$0.008_{\pm.00}$	$0.000_{\pm.00}$	$0.491_{\pm.01}$	$0.095_{\pm.01}$	$0.008_{\pm.00}$	$0.001_{\pm.00}$
Min-K%++	$0.494_{\pm.01}$	$0.109_{\pm.01}$	$0.011_{\pm.00}$	$0.000_{\pm.00}$	$0.488_{\pm.01}$	$0.118_{\pm.02}$	$0.004_{\pm.00}$	$0.000_{\pm.00}$	$0.491_{\pm.01}$	$0.113_{\pm.01}$	$0.007_{\pm.00}$	$0.001_{\pm.00}$
BoWs	$0.489_{\pm.01}$	$0.100_{\pm.01}$	$0.002_{\pm.00}$	$0.000_{\pm.00}$	$0.471_{\pm.01}$	$0.084_{\pm.02}$	$0.003_{\pm.00}$	$0.001_{\pm.00}$	$0.480_{\pm.01}$	$0.092_{\pm.01}$	$0.003_{\pm.00}$	$0.001_{\pm.00}$
ReCall	$0.495_{\pm.01}$	$0.088_{\pm.01}$	$0.007_{\pm.00}$	$0.000_{\pm.00}$	$0.506_{\pm.01}$	$0.103_{\pm.01}$	$0.010_{\pm.01}$	$0.000_{\pm.00}$	$0.500_{\pm.01}$	$0.096_{\pm.01}$	$0.009_{\pm.00}$	$0.000_{\pm.00}$
CON-Recall	$0.498_{\pm.02}$	$0.119_{\pm.02}$	$0.009_{\pm.00}$	$0.000_{\pm.00}$	$0.498_{\pm.02}$	$0.092_{\pm.01}$	$0.008_{\pm.00}$	$0.000_{\pm.00}$	$0.498_{\pm.01}$	$0.106_{\pm.01}$	$0.009_{\pm.00}$	$0.000_{\pm.00}$
Neighbor	$0.506_{\pm.01}$	$0.091_{\pm.01}$	$0.011_{\pm.01}$	$0.000_{\pm.00}$	$0.504_{\pm.01}$	$0.101_{\pm.01}$	$0.009_{\pm.00}$	$0.001_{\pm.00}$	$0.505_{\pm.01}$	$0.096_{\pm.01}$	$0.010_{\pm.00}$	$0.001_{\pm.00}$
Ratio	$0.555_{\pm.01}$	$0.128_{\pm.02}$	$0.018_{\pm.01}$	$0.003_{\pm.01}$	$0.653_{\pm.01}$	$0.184_{\pm.03}$	$0.011_{\pm.01}$	$0.000_{\pm.00}$	$0.604_{\pm.01}$	$0.156_{\pm.02}$	$0.015_{\pm.01}$	$0.002_{\pm.00}$
SecMI	$0.510_{\pm.01}$	$0.109_{\pm.01}$	$0.009_{\pm.00}$	$0.008_{\pm.00}$	$0.522_{\pm.01}$	$0.101_{\pm.01}$	$0.015_{\pm.00}$	$0.009_{\pm.00}$	$0.515_{\pm.01}$	$0.108_{\pm.01}$	$0.010_{\pm.00}$	$0.001_{\pm.00}$
PIA	$0.496_{\pm.01}$	$0.102_{\pm.01}$	$0.005_{\pm.00}$	$0.000_{\pm.00}$	$0.522_{\pm.01}$	$0.120_{\pm.01}$	$0.011_{\pm.00}$	$0.001_{\pm.00}$	$0.509_{\pm.01}$	$0.103_{\pm.01}$	$0.009_{\pm.00}$	$0.000_{\pm.00}$
SAMA (Ours)	$\mathbf{0.814}_{\pm.01}$	$\mathbf{0.442}_{\pm.03}$	$\mathbf{0.132}_{\pm.03}$	$\mathbf{0.011}_{\pm.01}$	$\mathbf{0.790}_{\pm.01}$	$\mathbf{0.433}_{\pm.03}$	$\mathbf{0.136}_{\pm.01}$	$\mathbf{0.008}_{\pm.02}$	$\mathbf{0.778}_{\pm.01}$	$\mathbf{0.408}_{\pm.03}$	$\mathbf{0.115}_{\pm.02}$	$\mathbf{0.009}_{\pm.01}$

- **TL;DR**
  - **SOTA:** Outperforms 12 ARM and diffusion baselines.
  - **Ablation:** Every SAMA component is necessary.
  - **Robustness:** SAMA is highly stable across hyperparameter changes.





Membership Inference Attacks Against Fine-tuned Diffusion Language Models

Yuetian Chen[1], Kaiyuan Zhang[1], Yuntao Du[1], Edoardo Stoppa[1], Charles Fleming[2], Ashish Kundu[2], Bruno Ribeiro[1], Ninghui Li[1]

[1] Department of Computer Science, Purdue University;

[2] Cisco Research

Thank you!

Paper Code Blog