Window-based Membership Inference Attacks Against Fine-tuned Large Language Models

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

[1] Department of Computer Science, Purdue University;

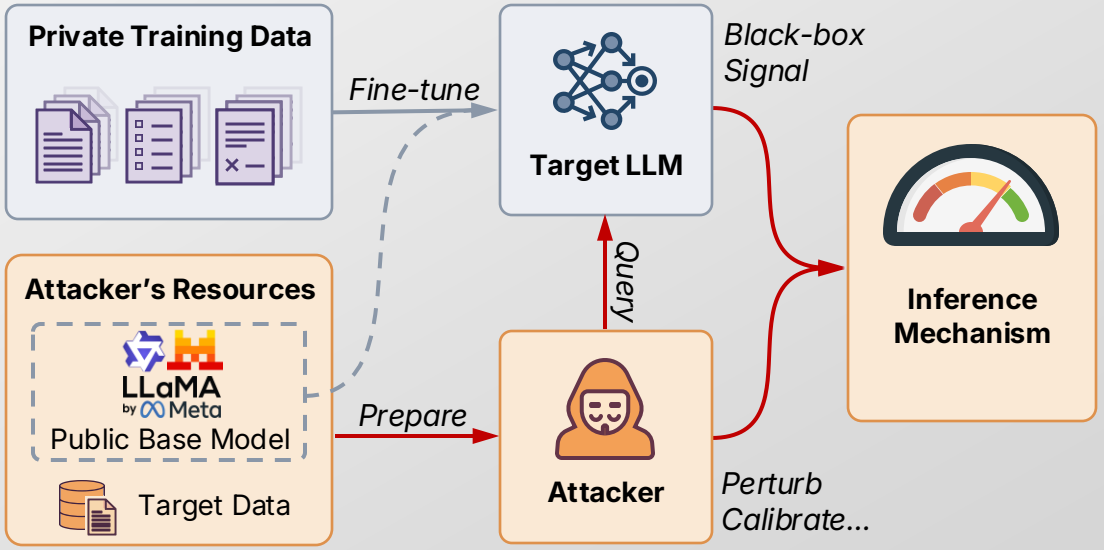
[2] Cisco Research

Presenter: Yuetian Chen

The Privacy Risk in Fine-Tuned LLMs

Was my data point used to train a model?

- **The Problem**
  - LLMs are fine-tuned on private datasets
    - *Medical records, legal documents, proprietary code…*
  - Fine-tuning causes the model to memorize training data
- **Membership Inference Attacks (MIA)**
  - Goal: Given text x and model, determine if x was in training data
- **Threat model of this work**
  - *Grey-box access*: query models, get per-token losses
  - Target model: fine-tuned model
  - Reference model: pre-trained base model
  - No access to parameters, gradients, or activations



The high-level pipeline of LLM MIA @ fine-tune scenario

Background: How Do Current LLM MIAs Work?

And are they correct?

- **The (High-level) Standard Approach**
  - Compute per-token loss for each model:

$$\ell_T^i = -\log p(x_i \mid x_1, \ldots, x_{i-1})$$

$$\ell_R^i = -\log p(x_i \mid x_1, \ldots, x_{i-1})$$

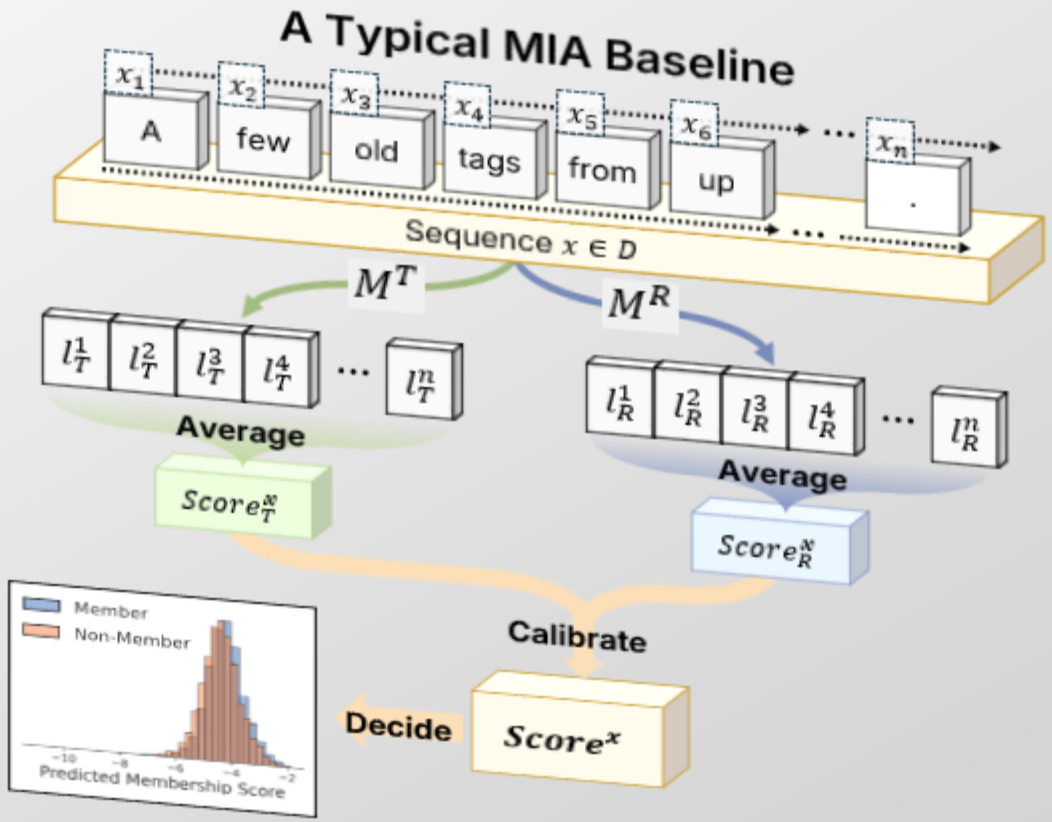
  - Aggregate over the ENTIRE* document

$$Score_{\{T,R\}}^x = \frac{1}{n}\sum_{i=1}^{n} \ell_{\{T,R\}}^i$$

  - Calibration – cancel out factors like sample difficulty

$$Score^x = \frac{Score_T^x}{Score_R^x}$$

  - Check how likely $Score^x$ is located at member or non-member distribution and make a decision



- **Underlying Assumption**
  - If model is *on average* more confident about this document, then it is more likely in the training data
  - **But is this assumption actually correct?**

**: also could be just part of it under some regularization*

Background: How Do Current LLM MIAs Work?

And are they correct?

- **The (High-level) Standard Approach**
  - Compute per-token loss for each model:

$$\ell_T^i = -\log p(x_i \mid x_1, \ldots, x_{i-1})$$

$$\ell_R^i = -\log p(x_i \mid x_1, \ldots, x_{i-1})$$

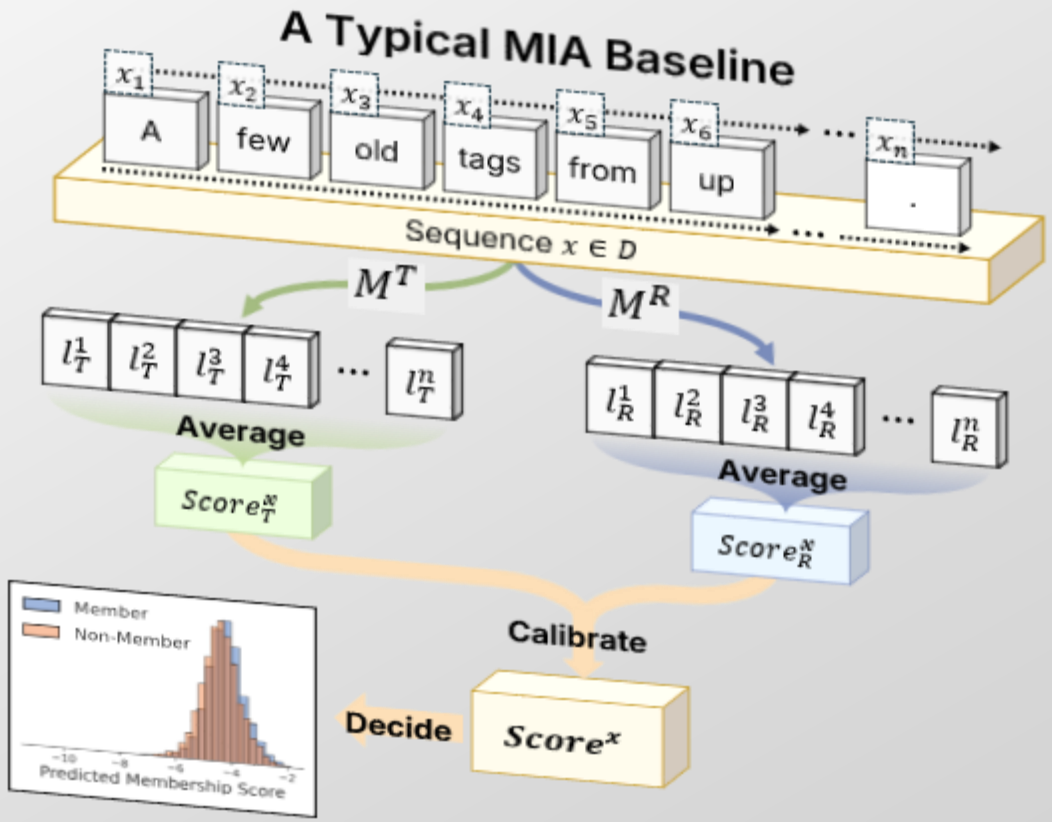
  - Aggregate over the ENTIRE* document

$$Score_{\{T,R\}}^x = \frac{1}{n}\sum_{i=1}^{n} \ell_{\{T,R\}}^i$$

  - Calibration – cancel out factors like sample difficulty

$$Score^x = \frac{Score_T^x}{Score_R^x}$$

  - Check how likely $Score^x$ is located at member or non-member distribution and make a decision



- **Underlying Assumption**
  - If model is *on average* more confident about this document, then it is more likely in the training data
  - **But is this assumption actually correct? – Not sharp enough.**

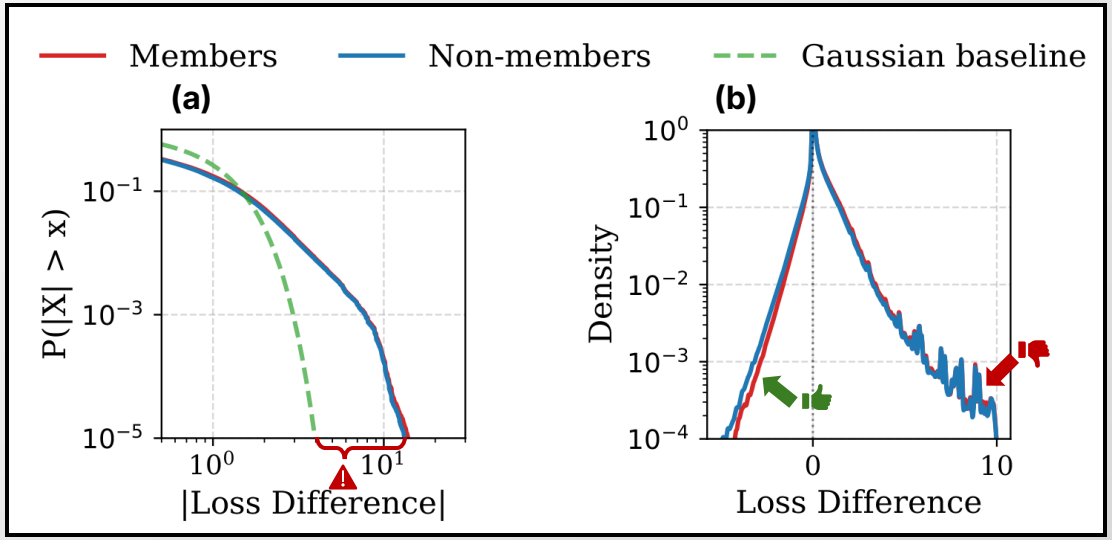
**: also could be just part of it under some regularization*

What Do Membership Signals Actually Look Like?

10M+ token-level loss differences @ Pythia-2.8B fine-tuned on Khan Academy

- **Panel (a): Extremely Long Tails**
  - Kurtosis > 18 (Gaussian = 0)
  - 1.8% of tokens exceed 3σ (Gaussian: 0.3%)
  - Individual extremes can be 10–100× larger than typical values

- **Panel (b): Two Tails, One Mechanism**
  - Right tail
    - Domain tokens receive enormous loss reductions
    - These events are rare and loud
      - but *not* membership-specific
  - Left tail
    - The displaced probability increases loss on many other tokens
    - Memorization partially protects member contexts from this increase
    - These effects are smaller, but more frequent than the rare right-tail spikes



- **Averaging is unreliable** when a single extreme domain token can dominate the entire sequence

- **Domain adaptation is loud and helpless**; membership appears as a smaller but more frequent left-tail shift.

Modeling Membership Signals as Extremal Events

Formalizing with mixture of point processes

- **Per-token loss difference decomposition**

$$\Delta_j(\mathbf{x}) = \ell_j^{\mathrm{R}} - \ell_j^{\mathrm{T}} = \underbrace{\mathbb{I}[\mathbf{x} \in D_{\mathrm{train}}] \cdot \delta_j(\mathbf{x})}_{\text{membership signal}} + \underbrace{\xi_j}_{\text{domain-specific noise}} + \underbrace{\epsilon_j}_{\text{baseline noise}}$$

- **Three Components**
  - ϵ_j : baseline noise
    - Small mean, finite variance, generic prediction fluctuations
  - ξ_j : domain-specific tokens *(the INTERFERENCE)*
    - Rare ($\rho \approx 0.02$) but EXTREME magnitude
    - Long-tailed: can be 10-100× typical values
    - Appears equally in members & non-members
    - **These dominate global averages**
  - δ_j : membership signal *(the TARGET)*
    - Also sparse, moderate magnitude
    - Only present for training members
    - Smaller than ξ but more discriminative

- **Why Global Averaging Fails (Formally)**
  - Global average decomposes as:

$$\bar{\Delta} = \frac{1}{n}\sum_{j=1}^{n} \Delta_j = \underbrace{\mathbb{I}[\mathbf{x} \in D_{\mathrm{train}}] \cdot \rho_\delta \bar{\gamma}}_{\text{membership signal}} + \underbrace{\frac{1}{n}\sum_{j=1}^{n} \xi_j}_{\text{rare token noise}} + \underbrace{\bar{\epsilon}}_{\text{baseline noise}}.$$

  - Signal: $O(\rho\delta \cdot \bar{\gamma})$ — small, fixed
  - Noise variance $\propto \rho\xi \cdot \mathrm{E}[Y^2]$ — can be infinite!
  - A single extreme ξ_j can completely dominate the entire membership score

The Key Insight

- Don’t average globally.
- Localize the impact of extremal events.

WBC: Window-Based Detection with Sign Aggregation

Replacing global averaging with local binary voting

- **Step 1: Sliding Window Comparison**
  - Reframe MIA: localize the impact of extremal events

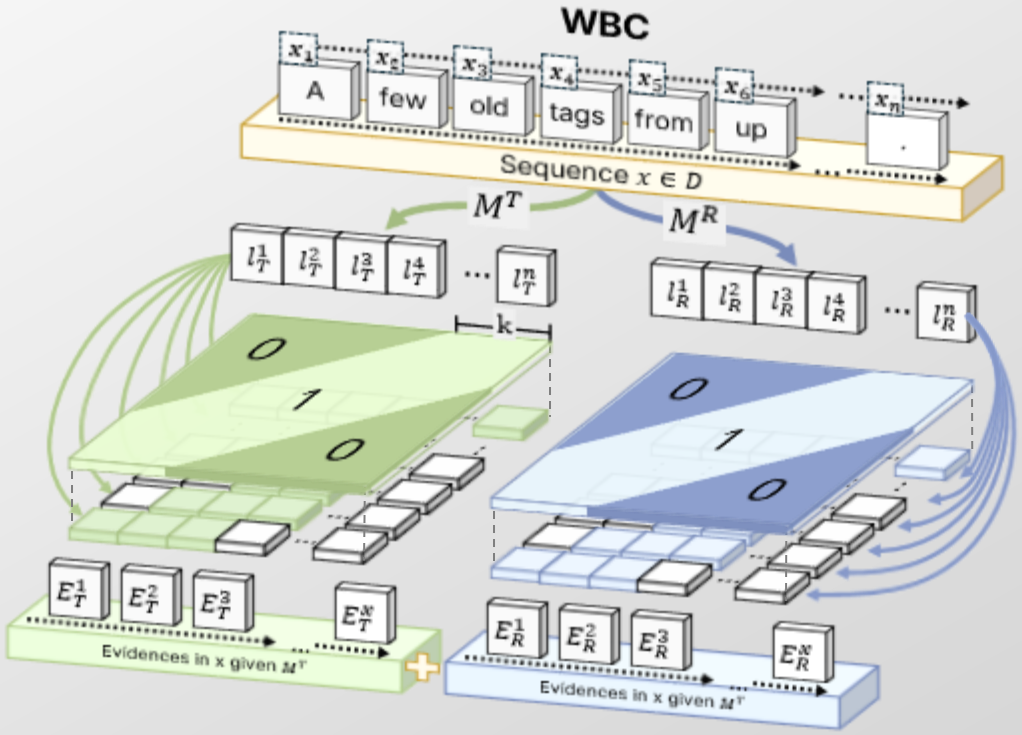
$$S_i(w) = \sum_{j=i}^{i+w-1} \Delta_j$$

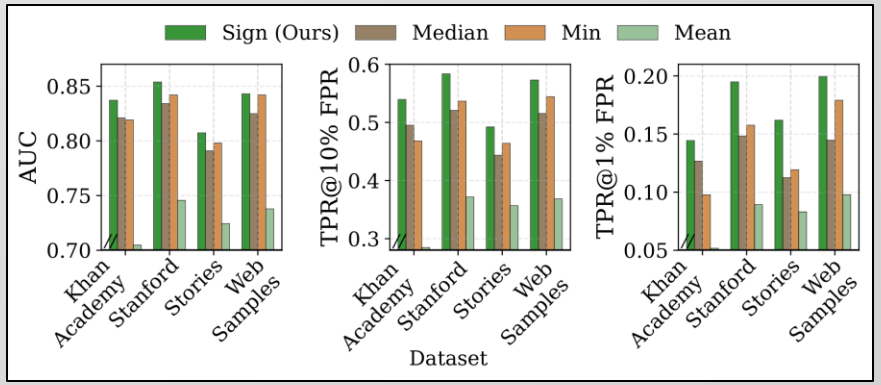
  - Slide across full sequence → (n−w+1) local tests

- **Step 2: Sign-Based (Binary) Aggregation**

$$T_{\text{sign}} = \frac{1}{n-w+1} \sum_{i=1}^{n-w+1} \mathbb{I}[S_i(w) > 0].$$

- **Why Sign? (Pitman ARE Theory)**
  - $ARE(sign, mean) = 4f^2(0) \cdot Var[S_i]$
    - Long-tailed → Var is huge, f(0) bounded → ARE > 1
    - ARE = 2–5 at observed contamination levels
    - Under Cauchy contamination: ARE → ∞
  - Breakdown point = 0.5 (maximum possible)
    - 50% contaminated windows can't break it
  - **"How OFTEN is loss lower" beats "How MUCH lower"**





WBC: Multi-Scale Geometric Ensemble

Eliminating the window size parameter entirely

- **The Window Size Dilemma**
  - Too small ($w = 1,2$):
    - High variance per test, poor signal-to-noise ratio
  - Too large ($w \gg 10$):
    - Signal dilution, outlier contamination
    - Fewer independent tests ($n_{eff} \approx n/w$)
    - Optimal w depends on unknown, varying parameters

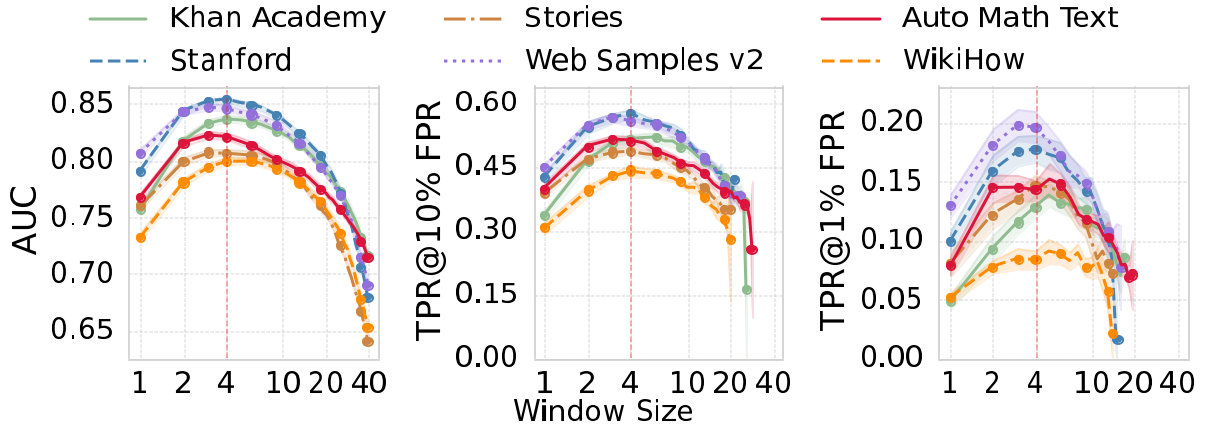
- **Solution: Geometric Window Progression**

$$w_k = round\left(w_{min} \cdot \left(\frac{w_{max}}{w_{min}}\right)^{\frac{k-1}{|W|-1}}\right)$$

  - Dense at small w (where signals concentrate)
  - Sparse at large w (coverage without redundancy)

Reminder - *Per-token loss difference decomposition*

$$\Delta_j(\mathbf{x}) = \ell_j^{\mathrm{R}} - \ell_j^{\mathrm{T}} = \underbrace{[\mathbf{x} \in D_{\mathrm{train}}] \cdot \delta_j(\mathbf{x})}_{\text{membership signal}} + \underbrace{\xi_j}_{\text{domain-specific noise}} + \underbrace{\epsilon_j}_{\text{baseline noise}}$$



Configuration	AUC	TPR@10%	TPR@1%
Single Best (w=4)	$0.8341_{\pm.0024}$	$0.5103_{\pm.0105}$	$0.1262_{\pm.0094}$
Small Range	$0.8340_{\pm.0028}$	$0.5195_{\pm.0091}$	$0.1296_{\pm.0090}$
Large Range	$0.7836_{\pm.0030}$	$0.4293_{\pm.0079}$	$0.0011_{\pm.0110}$
Full Ensemble	$0.8369_{\pm.0027}$	$0.5384_{\pm.0076}$	$0.1464_{\pm.0088}$
Extended	$0.8268_{\pm.0030}$	$0.5209_{\pm.0084}$	$0.1378_{\pm.0085}$
Linear Spacing	$0.8241_{\pm.0034}$	$0.5147_{\pm.0096}$	$0.1389_{\pm.0087}$
Random Selection	$0.8065_{\pm.0026}$	$0.4720_{\pm.0090}$	$0.1325_{\pm.0078}$

Evaluation: WBC vs. 13 Baselines on 11 Datasets

Pythia-2.8B @ 6 Cosmopedia + 5 real-world datasets

- **TL;DR**
  - WBC works!

Model	Method	Attack Performance			
		AUC	TPR@10%	TPR@1%	TPR@0.1%
GPT-2 [83]	Ratio	$0.505_{\pm.004}$	$0.101_{\pm.003}$	$0.009_{\pm.001}$	$0.001_{\pm.001}$
	WBC	$\mathbf{0.505}_{\pm.003}$	$\mathbf{0.104}_{\pm.003}$	$\mathbf{0.013}_{\pm.002}$	$\mathbf{0.001}_{\pm.001}$
Mamba-1.4B [32]	Ratio	$0.592_{\pm.003}$	$0.154_{\pm.005}$	$0.018_{\pm.002}$	$0.001_{\pm.000}$
	WBC	$\mathbf{0.673}_{\pm.002}$	$\mathbf{0.274}_{\pm.004}$	$\mathbf{0.048}_{\pm.003}$	$\mathbf{0.004}_{\pm.001}$
Pythia-2.8B [6]	Ratio	$0.703_{\pm.003}$	$0.264_{\pm.004}$	$0.037_{\pm.004}$	$0.003_{\pm.001}$
	WBC	$\mathbf{0.833}_{\pm.002}$	$\mathbf{0.529}_{\pm.009}$	$\mathbf{0.135}_{\pm.008}$	$\mathbf{0.026}_{\pm.009}$
Llama-3.2-3B [25]	Ratio	$0.861_{\pm.003}$	$0.551_{\pm.010}$	$0.135_{\pm.014}$	$0.019_{\pm.006}$
	WBC	$\mathbf{0.942}_{\pm.002}$	$\mathbf{0.831}_{\pm.006}$	$\mathbf{0.467}_{\pm.015}$	$\mathbf{0.193}_{\pm.028}$
GPT-J-6B [98]	Ratio	$0.868_{\pm.002}$	$0.564_{\pm.009}$	$0.126_{\pm.007}$	$0.013_{\pm.005}$
	WBC	$\mathbf{0.961}_{\pm.001}$	$\mathbf{0.887}_{\pm.003}$	$\mathbf{0.566}_{\pm.016}$	$\mathbf{0.209}_{\pm.043}$

Configuration	AUC	Δ	TPR@10%	TPR@1%
Full WBC	$\mathbf{0.8369}_{\pm.0027}$	—	$\mathbf{0.5384}_{\pm.0076}$	$\mathbf{0.1464}_{\pm.0088}$
Design choices				
w/o ensemble	$0.8341_{\pm.0026}$	−0.3%	$0.5093_{\pm.0092}$	$0.1278_{\pm.0091}$
w/o geometric	$0.8152_{\pm.0029}$	−2.6%	$0.5023_{\pm.0088}$	$0.1373_{\pm.0080}$
w/o sliding	$0.8042_{\pm.0030}$	−3.9%	$0.4195_{\pm.0062}$	$0.0715_{\pm.0099}$
Core components				
w/o sign-based	$0.7045_{\pm.0038}$	−15.8%	$0.2835_{\pm.0077}$	$0.0508_{\pm.0046}$
w/o reference	$0.5693_{\pm.0040}$	−31.9%	$0.1510_{\pm.0053}$	$0.0175_{\pm.0022}$

MIAs	Khan Academy				Stanford				Stories			
	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 [107]	$0.568_{\pm.003}$	$0.151_{\pm.003}$	$0.017_{\pm.002}$	$0.001_{\pm.001}$	$0.590_{\pm.006}$	$0.170_{\pm.008}$	$0.023_{\pm.003}$	$0.003_{\pm.001}$	$0.587_{\pm.004}$	$0.146_{\pm.008}$	$0.020_{\pm.001}$	$0.003_{\pm.001}$
ZLIB [9]	$0.583_{\pm.004}$	$0.161_{\pm.003}$	$0.019_{\pm.003}$	$0.003_{\pm.001}$	$0.606_{\pm.006}$	$0.178_{\pm.009}$	$0.024_{\pm.003}$	$0.002_{\pm.000}$	$0.596_{\pm.004}$	$0.157_{\pm.005}$	$0.019_{\pm.002}$	$0.003_{\pm.001}$
Lowercase [9]	$0.586_{\pm.004}$	$0.154_{\pm.006}$	$0.013_{\pm.002}$	$0.002_{\pm.001}$	$0.596_{\pm.003}$	$0.173_{\pm.004}$	$0.024_{\pm.002}$	$0.003_{\pm.001}$	$0.593_{\pm.004}$	$0.172_{\pm.004}$	$0.018_{\pm.002}$	$0.002_{\pm.001}$
Min-K% [91]	$0.595_{\pm.002}$	$0.182_{\pm.004}$	$0.024_{\pm.002}$	$0.002_{\pm.000}$	$0.593_{\pm.005}$	$0.168_{\pm.006}$	$0.024_{\pm.002}$	$0.003_{\pm.001}$	$0.588_{\pm.003}$	$0.145_{\pm.004}$	$0.019_{\pm.003}$	$0.003_{\pm.001}$
Min-K%++ [109]	$0.596_{\pm.002}$	$0.180_{\pm.005}$	$0.022_{\pm.003}$	$0.002_{\pm.001}$	$0.592_{\pm.004}$	$0.161_{\pm.006}$	$0.023_{\pm.002}$	$0.004_{\pm.001}$	$0.586_{\pm.003}$	$0.150_{\pm.006}$	$0.018_{\pm.002}$	$0.002_{\pm.001}$
BoWs [16]	$0.499_{\pm.004}$	$0.101_{\pm.003}$	$0.011_{\pm.001}$	$0.001_{\pm.001}$	$0.498_{\pm.003}$	$0.109_{\pm.003}$	$0.012_{\pm.001}$	$0.001_{\pm.001}$	$0.502_{\pm.005}$	$0.100_{\pm.002}$	$0.011_{\pm.001}$	$0.002_{\pm.001}$
ReCall [104]	$0.568_{\pm.003}$	$0.151_{\pm.003}$	$0.017_{\pm.002}$	$0.001_{\pm.001}$	$0.590_{\pm.006}$	$0.170_{\pm.008}$	$0.023_{\pm.003}$	$0.003_{\pm.001}$	$0.587_{\pm.004}$	$0.146_{\pm.008}$	$0.020_{\pm.001}$	$0.003_{\pm.001}$
CON-Recall [99]	$0.563_{\pm.003}$	$0.142_{\pm.004}$	$0.014_{\pm.002}$	$0.002_{\pm.001}$	$0.570_{\pm.006}$	$0.159_{\pm.007}$	$0.018_{\pm.002}$	$0.003_{\pm.000}$	$0.558_{\pm.004}$	$0.143_{\pm.006}$	$0.020_{\pm.002}$	$0.002_{\pm.001}$
DC-PDD [113]	$0.567_{\pm.003}$	$0.136_{\pm.003}$	$0.015_{\pm.002}$	$0.002_{\pm.001}$	$0.570_{\pm.003}$	$0.140_{\pm.003}$	$0.014_{\pm.002}$	$0.002_{\pm.001}$	$0.574_{\pm.002}$	$0.146_{\pm.002}$	$0.019_{\pm.002}$	$0.002_{\pm.001}$
SPV-MIA [29]	$0.695_{\pm.003}$	$0.240_{\pm.006}$	$0.049_{\pm.004}$	$0.005_{\pm.003}$	$0.760_{\pm.003}$	$0.260_{\pm.010}$	$0.077_{\pm.005}$	$0.012_{\pm.003}$	$0.763_{\pm.004}$	$0.360_{\pm.007}$	$0.070_{\pm.006}$	$0.018_{\pm.005}$
Ratio [100]	$0.703_{\pm.003}$	$0.264_{\pm.004}$	$0.037_{\pm.004}$	$0.003_{\pm.001}$	$0.781_{\pm.004}$	$0.401_{\pm.010}$	$0.080_{\pm.008}$	$0.013_{\pm.001}$	$0.769_{\pm.004}$	$0.389_{\pm.011}$	$0.091_{\pm.008}$	$0.022_{\pm.006}$
Difference [100]	$0.692_{\pm.002}$	$0.259_{\pm.005}$	$0.045_{\pm.002}$	$0.003_{\pm.001}$	$0.742_{\pm.005}$	$0.360_{\pm.011}$	$0.080_{\pm.008}$	$0.021_{\pm.004}$	$0.719_{\pm.005}$	$0.348_{\pm.007}$	$0.079_{\pm.006}$	$0.014_{\pm.003}$
Ensemble [110]	$0.687_{\pm.003}$	$0.217_{\pm.004}$	$0.061_{\pm.002}$	$0.007_{\pm.001}$	$0.738_{\pm.004}$	$0.142_{\pm.003}$	$0.075_{\pm.006}$	$0.011_{\pm.004}$	$0.758_{\pm.004}$	$0.338_{\pm.004}$	$0.048_{\pm.003}$	$0.014_{\pm.003}$
WBC (Ours)	$\mathbf{0.837}_{\pm.003}$	$\mathbf{0.538}_{\pm.007}$	$\mathbf{0.146}_{\pm.009}$	$\mathbf{0.026}_{\pm.009}$	$\mathbf{0.854}_{\pm.003}$	$\mathbf{0.583}_{\pm.008}$	$\mathbf{0.194}_{\pm.012}$	$\mathbf{0.034}_{\pm.017}$	$\mathbf{0.808}_{\pm.005}$	$\mathbf{0.494}_{\pm.007}$	$\mathbf{0.160}_{\pm.013}$	$\mathbf{0.034}_{\pm.003}$

MIAs	Web Samples v2				Auto Math Text				WikiHow			
	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 [107]	$0.579_{\pm.004}$	$0.153_{\pm.006}$	$0.022_{\pm.003}$	$0.004_{\pm.001}$	$0.553_{\pm.005}$	$0.133_{\pm.005}$	$0.014_{\pm.001}$	$0.002_{\pm.001}$	$0.592_{\pm.004}$	$0.162_{\pm.005}$	$0.022_{\pm.002}$	$0.004_{\pm.001}$
ZLIB [9]	$0.593_{\pm.004}$	$0.175_{\pm.005}$	$0.023_{\pm.002}$	$0.005_{\pm.001}$	$0.562_{\pm.005}$	$0.141_{\pm.005}$	$0.015_{\pm.002}$	$0.002_{\pm.001}$	$0.604_{\pm.004}$	$0.171_{\pm.004}$	$0.022_{\pm.002}$	$0.002_{\pm.000}$
Lowercase [9]	$0.592_{\pm.004}$	$0.166_{\pm.005}$	$0.021_{\pm.003}$	$0.003_{\pm.001}$	$0.568_{\pm.003}$	$0.141_{\pm.008}$	$0.015_{\pm.002}$	$0.001_{\pm.001}$	$0.614_{\pm.004}$	$0.182_{\pm.007}$	$0.024_{\pm.002}$	$0.001_{\pm.001}$
Min-K% [91]	$0.587_{\pm.005}$	$0.158_{\pm.007}$	$0.023_{\pm.002}$	$0.002_{\pm.000}$	$0.581_{\pm.006}$	$0.150_{\pm.006}$	$0.021_{\pm.001}$	$0.002_{\pm.000}$	$0.600_{\pm.004}$	$0.178_{\pm.005}$	$0.025_{\pm.002}$	$0.004_{\pm.001}$
Min-K%++ [109]	$0.586_{\pm.004}$	$0.153_{\pm.005}$	$0.020_{\pm.002}$	$0.003_{\pm.001}$	$0.583_{\pm.005}$	$0.150_{\pm.005}$	$0.022_{\pm.002}$	$0.003_{\pm.001}$	$0.598_{\pm.004}$	$0.174_{\pm.005}$	$0.027_{\pm.003}$	$0.004_{\pm.001}$
BoWs [16]	$0.493_{\pm.004}$	$0.097_{\pm.004}$	$0.010_{\pm.001}$	$0.002_{\pm.001}$	$0.500_{\pm.004}$	$0.105_{\pm.005}$	$0.009_{\pm.001}$	$0.001_{\pm.000}$	$0.498_{\pm.004}$	$0.096_{\pm.003}$	$0.008_{\pm.001}$	$0.001_{\pm.000}$
ReCall [104]	$0.579_{\pm.004}$	$0.153_{\pm.006}$	$0.022_{\pm.003}$	$0.004_{\pm.001}$	$0.553_{\pm.005}$	$0.133_{\pm.005}$	$0.014_{\pm.001}$	$0.002_{\pm.001}$	$0.592_{\pm.004}$	$0.162_{\pm.005}$	$0.022_{\pm.002}$	$0.004_{\pm.001}$
CON-Recall [99]	$0.560_{\pm.004}$	$0.144_{\pm.003}$	$0.017_{\pm.001}$	$0.002_{\pm.001}$	$0.547_{\pm.005}$	$0.135_{\pm.005}$	$0.015_{\pm.001}$	$0.002_{\pm.001}$	$0.571_{\pm.005}$	$0.161_{\pm.005}$	$0.019_{\pm.002}$	$0.002_{\pm.001}$
DC-PDD [113]	$0.569_{\pm.003}$	$0.143_{\pm.003}$	$0.019_{\pm.002}$	$0.003_{\pm.001}$	$0.560_{\pm.006}$	$0.130_{\pm.006}$	$0.015_{\pm.002}$	$0.002_{\pm.001}$	$0.571_{\pm.004}$	$0.143_{\pm.003}$	$0.016_{\pm.002}$	$0.002_{\pm.001}$
SPV-MIA [29]	$0.787_{\pm.003}$	$0.456_{\pm.006}$	$0.088_{\pm.004}$	$0.008_{\pm.006}$	$0.759_{\pm.003}$	$0.369_{\pm.006}$	$0.063_{\pm.006}$	$0.007_{\pm.004}$	$0.719_{\pm.003}$	$0.300_{\pm.007}$	$0.053_{\pm.005}$	$0.013_{\pm.004}$
Ratio [100]	$0.788_{\pm.003}$	$0.435_{\pm.009}$	$0.090_{\pm.007}$	$0.012_{\pm.003}$	$0.768_{\pm.002}$	$0.378_{\pm.005}$	$0.075_{\pm.002}$	$0.011_{\pm.006}$	$0.714_{\pm.003}$	$0.261_{\pm.006}$	$0.042_{\pm.003}$	$0.008_{\pm.002}$
Difference [100]	$0.739_{\pm.003}$	$0.365_{\pm.008}$	$0.094_{\pm.008}$	$0.019_{\pm.007}$	$0.700_{\pm.004}$	$0.300_{\pm.009}$	$0.051_{\pm.006}$	$0.009_{\pm.003}$	$0.709_{\pm.003}$	$0.278_{\pm.005}$	$0.047_{\pm.003}$	$0.009_{\pm.002}$
Ensemble [110]	$0.786_{\pm.003}$	$0.477_{\pm.004}$	$0.086_{\pm.002}$	$0.005_{\pm.013}$	$0.749_{\pm.002}$	$0.359_{\pm.004}$	$0.051_{\pm.002}$	$0.003_{\pm.002}$	$0.724_{\pm.003}$	$0.338_{\pm.004}$	$0.065_{\pm.005}$	$0.018_{\pm.001}$
WBC (Ours)	$\mathbf{0.843}_{\pm.003}$	$\mathbf{0.573}_{\pm.008}$	$\mathbf{0.198}_{\pm.012}$	$\mathbf{0.045}_{\pm.004}$	$\mathbf{0.814}_{\pm.003}$	$\mathbf{0.504}_{\pm.007}$	$\mathbf{0.153}_{\pm.009}$	$\mathbf{0.040}_{\pm.008}$	$\mathbf{0.802}_{\pm.003}$	$\mathbf{0.451}_{\pm.010}$	$\mathbf{0.096}_{\pm.007}$	$\mathbf{0.019}_{\pm.006}$

Take-away

- **Core Insight**
  - The largest confidence gain is usually domain adaptation, not memorization.
  - Membership appears as a repeated local advantage: small, scattered effects that consistently favor members across short regions.

- **WBC turns local consistency into a robust membership score:**
  - Slide windows over target–reference token losses
  - Convert each local comparison into a binary vote
  - Aggregate votes across multiple window sizes

- **Result:** 2–3× higher detection across 11 datasets and 13 baselines, with no additional model forward passes.

PURDUE UNIVERSITY CISCO USENIX

Paper Code Blog

Window-based Membership Inference Attacks Against Fine-tuned Large Language Models









Yuetian Chen, Yuntao Du, Kaiyuan Zhang, Ashish Kundu, Charles Fleming, Bruno Ribeiro, Ninghui Li

Thank you!