Membership Inference Attacks against Fine-tuned Large Language Models via Self-prompt Calibration

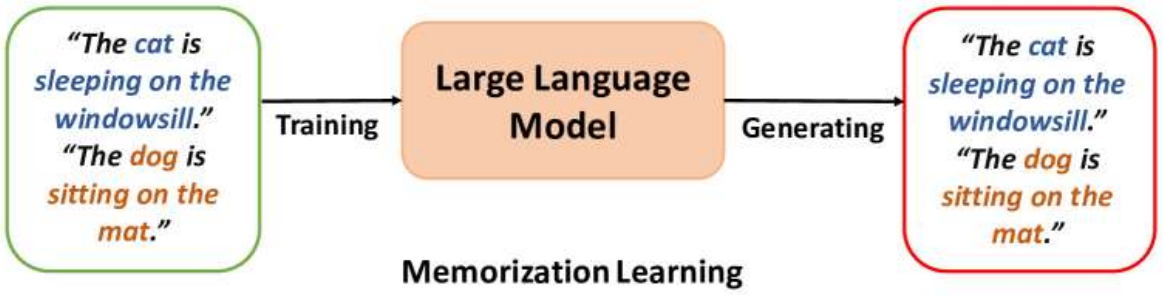
Wenjie Fu[1], Huandong Wang[2*], Chen Gao[2], Guanghua Liu[1], Yong Li[2], Tao Jiang[1]

[1] Huazhong University of Science and Technology, [2] Tsinghua University

CONTENTS

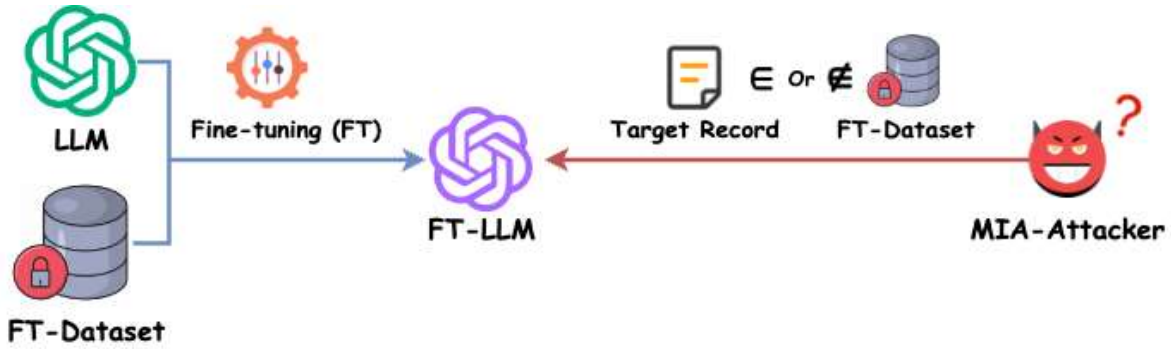
Background

- **Membership Inference Attack (MIA) against ML Models**
  - ✓ The training samples will be memorized by ML models



The training samples will be **memorized**.
(Tend to have lower loss)

  - ✓ Infer whether a given sample is included for training

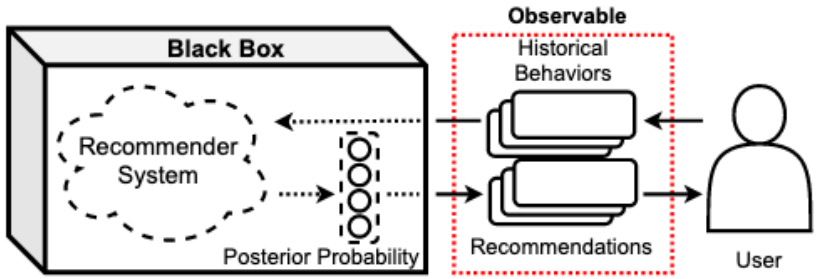


Whether a given data sample **is used to training**?

Member or **Non-Member**?

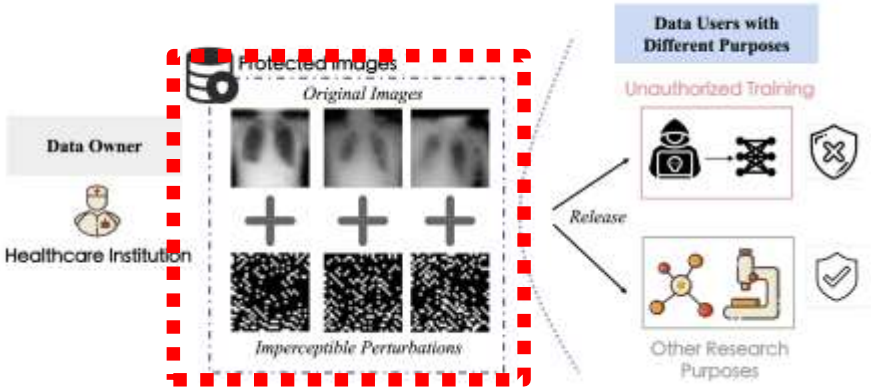
Background

- **Typical Applications**
  - ✓ Expose privacy via membership inference
    E.g., recommendation system



Whether a user had **used specific service**?

  - ✓ Detecting unauthorized content usage
    E.g., medical data, copyrighted works

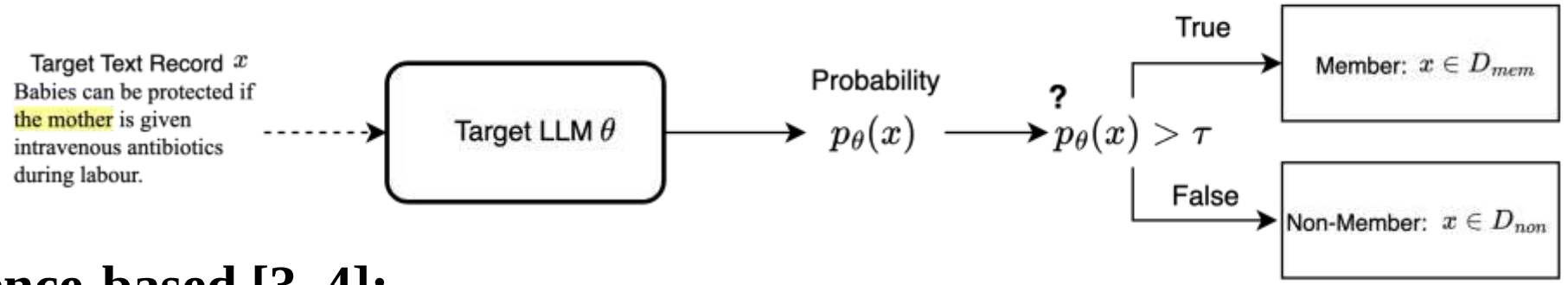


Whether **unauthorized data** is used for training?

Related Works: MIAs

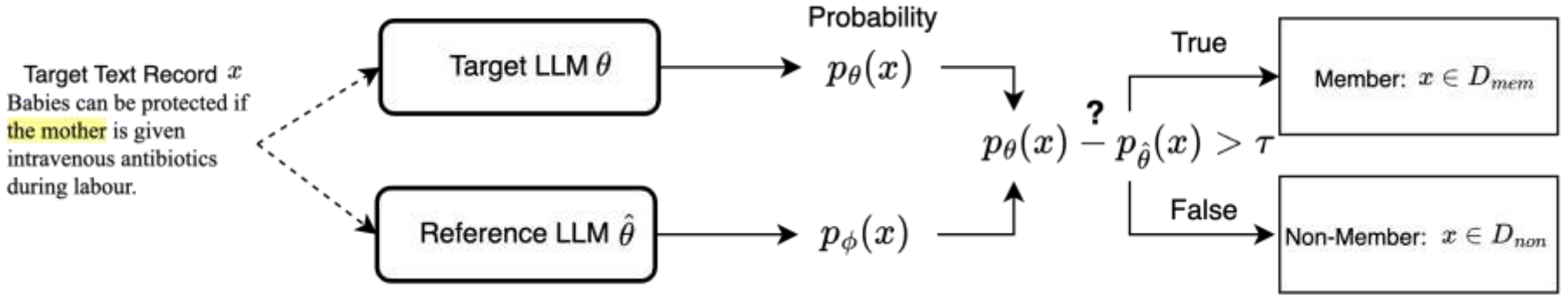
➢Reference-free [1, 2]:

- ✓ Only based on the target sample **probability (loss) of being generated** by the target language model
- ✓ The simplest method: taking the probability (loss) of target sample as the metric for MIA



➢Reference-based [3, 4]:

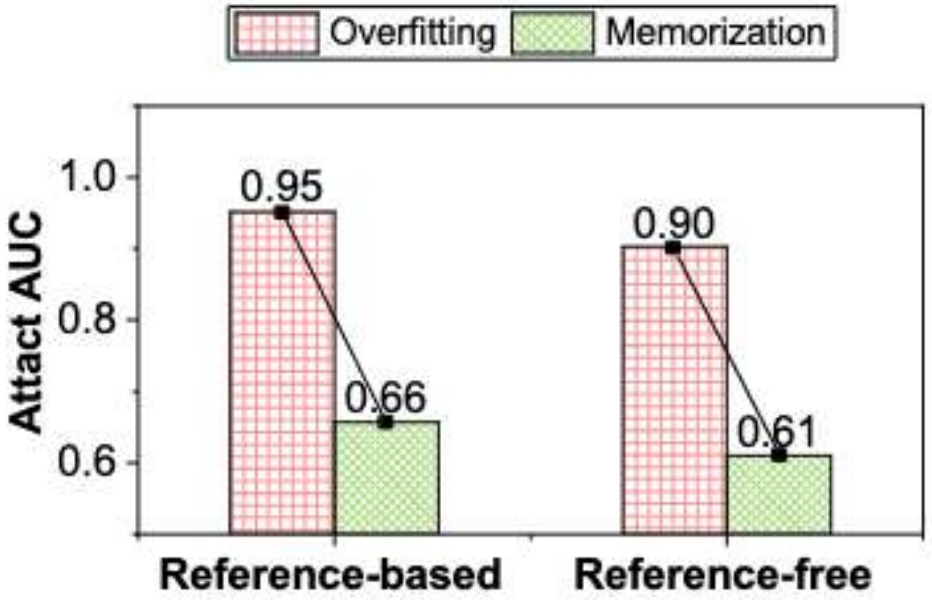
- ✓ Using a reference model to **calibrate the probability,** then select the abnormal high value

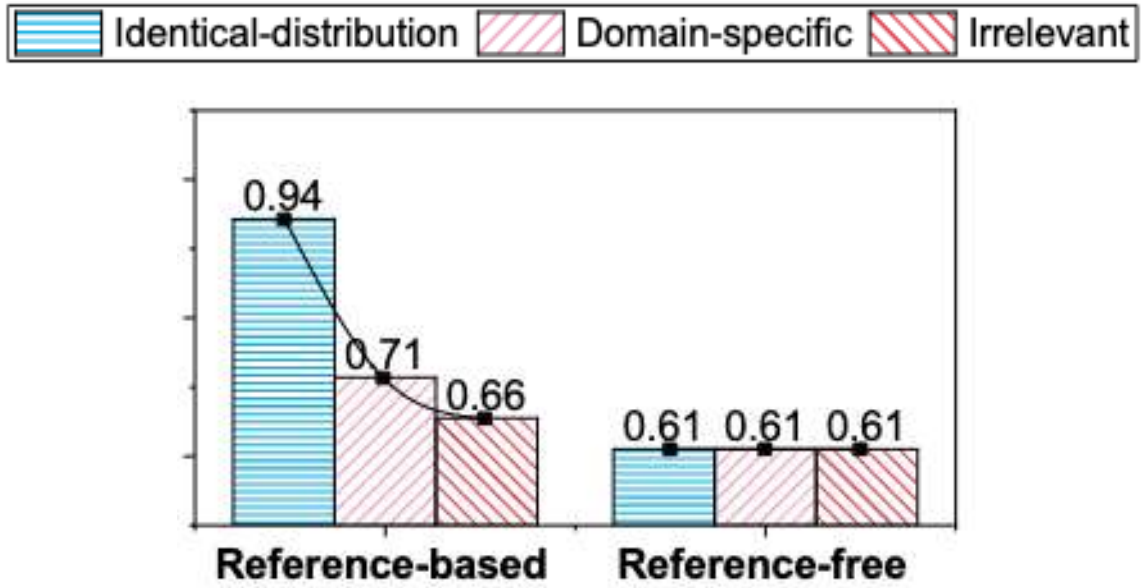


[1] Mattern, Justus, et al. "Membership Inference Attacks against Language Models via Neighbourhood Comparison." ACL'23
[2] Shi, Weijia, et al. "Detecting Pretraining Data from Large Language Models." ICLR'24
[3] Mireshghallah, Fatemehsadat, et al. "Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks." EMNLP'22
[4] Mireshghallah, Fatemehsadat, et al. "An Empirical Analysis of Memorization in Fine-Tuned Autoregressive Language Models." EMNLP'22

Related Works: MIAs

➢Limitations of Existing MIAs

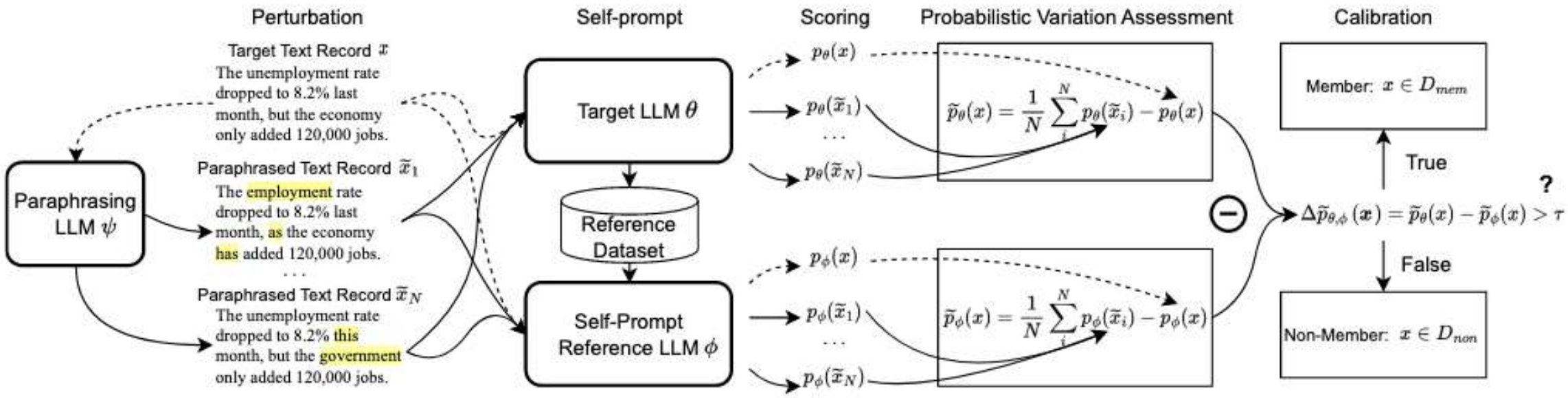




- ➢ **Bad Performance on the Practical Scenario**
- ✓ Only works on **overfitting LLMs** → can be easily avoided by regularization techniques
- ✓ Only works with **high quality reference dataset** → usually not accessible

Method: Overview

- **Membership Inference Attack based on Self-calibrated Probabilistic Variation (SPV-MIA).**

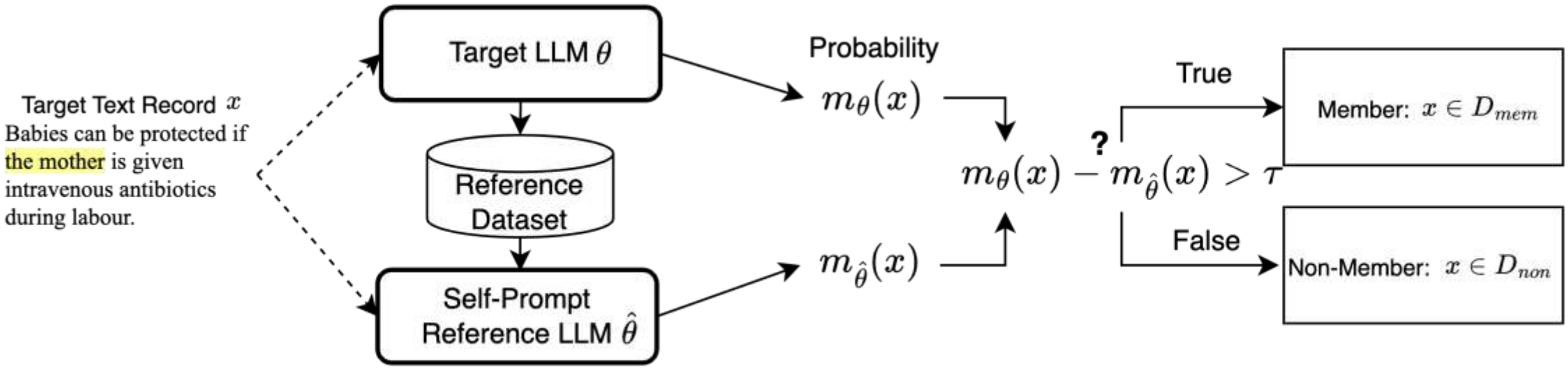


- Practical Difficulty Calibration (PDC) → Low quality of accessible reference datasets
- Probabilistic Variation Assessment (PVA) → Overfitting-free FT-LLMs

Method : Practical Difficulty Calibration

➢ Calibration via Self-Prompt Reference Model

✓ LLMs themselves may have the potential to generate **high quality reference dataset!**

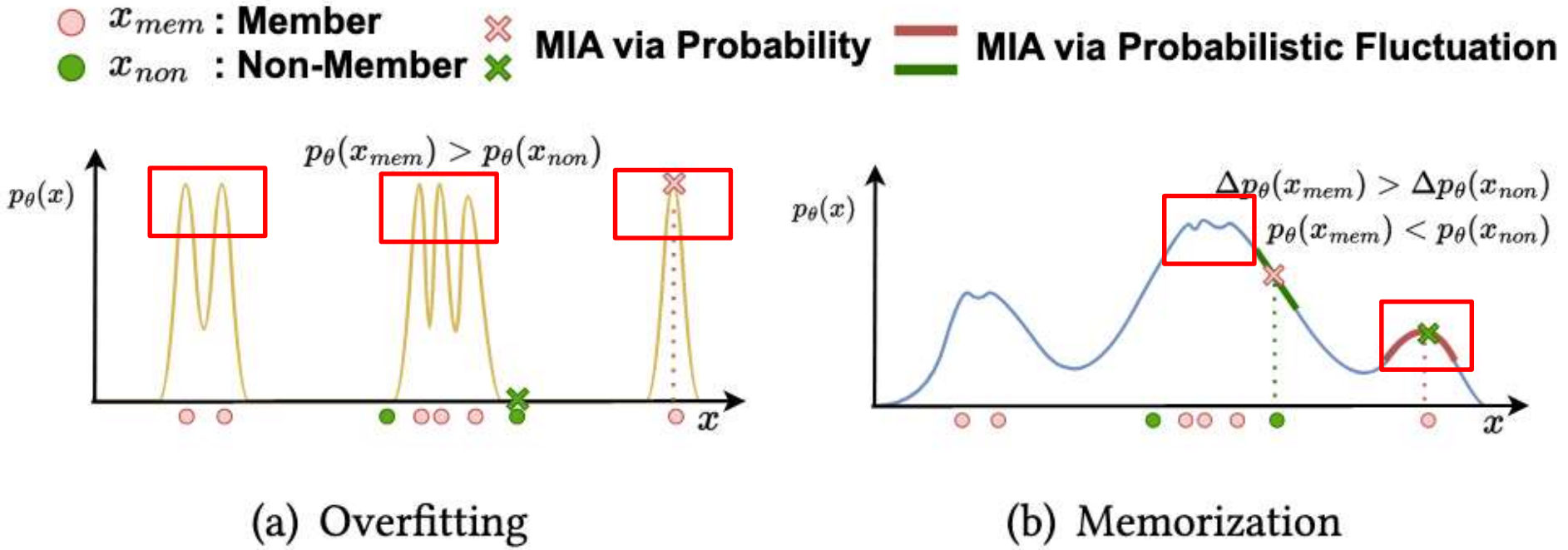


$$\Delta m\left(\boldsymbol{x}\right) = m_\theta\left(\boldsymbol{x}\right) - \mathbb{E}_{\widehat{\theta} \leftarrow \mathcal{T}\left(\mathcal{D}_{self}\right)}\left[m_{\widehat{\theta}}\left(\boldsymbol{x}\right)\right] \approx m_\theta\left(\boldsymbol{x}\right) - m_{\widehat{\theta}}\left(\boldsymbol{x}\right),$$

Method : Probabilistic Variation Assessment

➢ **Memorization rather than Overfitting**

✓ Memorization is a **more robust signal** for performing MIA**!**



(a) Overfitting (b) Memorization

$$\tilde{p}_\omega(x) \rightarrow \frac{1}{2N}\sum_{n}^{N}\left(p_\omega\left(\tilde{x}_n^{+}\right) + p_\omega\left(\tilde{x}_n^{\nearrow}\right)\right) \nearrow p_\omega(x)$$

Experiment: Overall Performance

➢ **Dose SPV-MIA outperform the state-of-the-art MIAs?**

Table 1: AUC for detecting member texts from four LLMs across three datasets for SPV-MIA and five previously proposed methods. Bold and Underline respectively represent the best and the second-best results within each column (model-dataset pair).

Method	Wiki					AG News					Xsum				
	GPT-2	GPT-J	Falcon	LLaMA	Avg.	GPT-2	GPT-J	Falcon	LLaMA	Avg.	GPT-2	GPT-J	Falcon	LLaMA	Avg.
Loss Attack	0.614	0.577	0.593	0.605	0.597	0.591	0.529	0.554	0.580	0.564	0.628	0.564	0.577	0.594	0.591
Neighbour Attack	0.647	0.612	0.621	0.627	0627	0.622	0.587	0.594	0.610	0.603	0.612	0.547	0.571	0.582	0.578
DetectGPT	0.623	0.587	0.603	0.619	0.608	0.611	0.579	0.582	0.603	0.594	0.603	0.541	0.563	0.577	0.571
LiRA-Base	0.710	0.681	0.694	0.709	0.699	0.658	0.634	0.641	0.657	0.648	0.776	0.718	0.734	0.759	0.747
LiRA-Candidate	0.769	0.726	0.735	0.748	0.744	0.717	0.690	0.708	0.714	0.707	0.823	0.772	0.785	0.809	0.797
Our	**0.975**	**0.929**	**0.932**	**0.951**	**0.938**	**0.949**	**0.885**	**0.898**	**0.903**	**0.909**	**0.944**	**0.897**	**0.918**	**0.937**	**0.924**

Full-training Fine-tuning

➢ **Conclusions**

- ✓ SPV-MIA consistently outperforms all baselines over all LLMs with different LLM architectures and fine-tuning datasets (AUC ~0.75 → ~0.92)
- ✓ The privacy risk caused by MIAs on LLMs is positively correlated with the overall NLP performance of the LLM itself

Thanks for your attention!