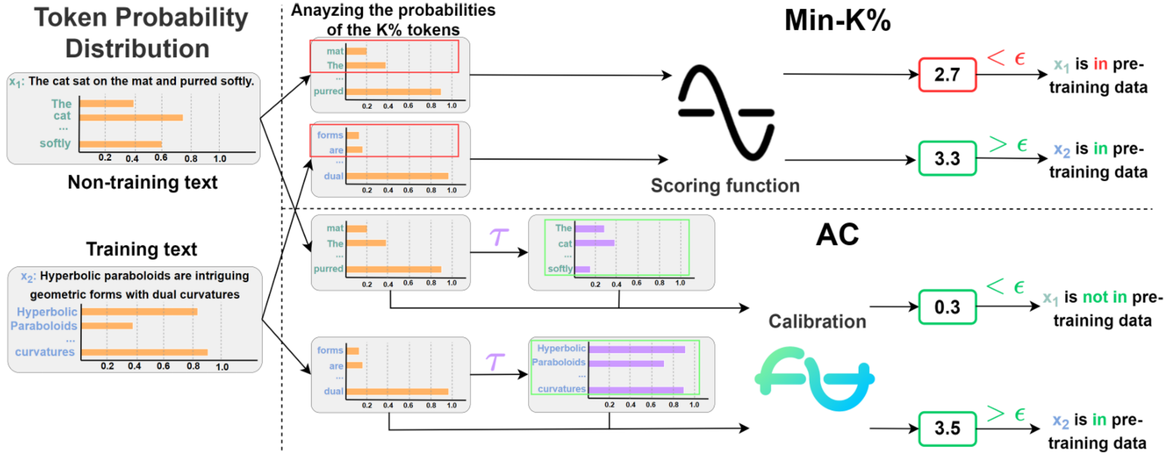
ACMIA: Automatic Calibration for Membership Inference Attack on Large Language Models

Saleh Zare Zade[1*], Yao Qiang[1*], Xiangyu Zhou[1], Hui Zhu[1], Mohammad Amin Roshani[1], Prashant Khanduri[1], Dongxiao Zhu[1]

[1]Wayne State University, *These authors contributed equally

Challenges: Membership Inference Attack (MIA) in LLMs



Motivation

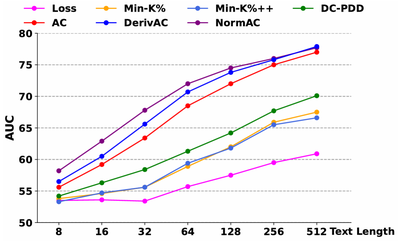
- LLMs memorize training data, raising concerns about privacy, data contamination, and copyright violations.
- Existing MIAs struggle with high false positive rates and often need reference models, which limit practical use.
- Simple or short texts naturally receive high probabilities from LLMs, even if they're not in the training set, making inference unreliable.

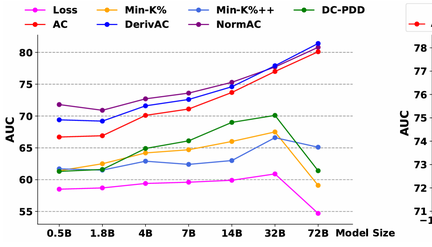
Our Solution: ACMIA

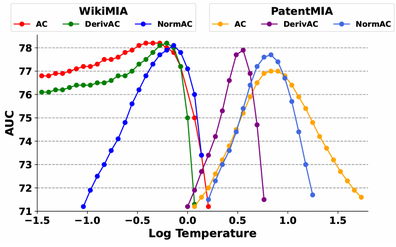
- ACMIA uses a tunable temperature to calibrate probabilities dynamically, removing the need for reference models.
- Includes three versions each designed based on different levels of access to model outputs.
- Temperature scaling adjusts the output distribution to amplify differences in the first- and second-order derivatives of the log-likelihood with respect to the input, helping reveal whether a sample lies near a local maximum.

Experimental Results

Method	PatentMIA				WikiMIA						MIMIR			
	Baichuan		Qwen1.5		OPT-6.7B		Pythia-12B		NeoX-20B		Pythia			
	13B	2-13B	32B	72B	*Ori.*	*Para.*	*Ori.*	*Para.*	*Ori.*	*Para.*	1.4B	2.8B	6.9B	12B
Loss	60.8	63.0	60.9	54.7	62.4	61.5	65.3	65.1	70.4	69.3	80.2	80.5	81.2	81.6
Ref	60.4	47.4	57.6	34.2	63.6	63.7	62.2	60.6	68.0	67.7	65.9	63.9	64.9	64.1
Lowercase	-	-	-	-	58.4	57.4	60.4	60.0	66.8	66.7	76.3	78.1	79.0	78.8
Zlib	63.6	67.6	63.6	53.5	64.3	64.3	67.5	67.7	72.0	71.8	77.4	77.9	78.6	78.9
Min-K%	66.7	70.1	67.5	59.1	67.4	64.7	69.8	67.8	75.5	72.3	80.6	80.9	81.7	82.2
Min-K%++	62.8	66.6	66.6	65.1	69.0	64.9	71.4	67.8	75.4	71.8	71.4	73.2	74.0	75.5
DC-PDD	70.0	74.7	70.1	61.4	67.9	66.1	70.1	68.1	75.8	73.1	81.4	81.7	82.3	82.5
AC	73.4	78.0	77.0	80.1	70.1	68.0	72.6	70.3	**78.2**	75.8	81.7	**82.3**	83.0	**83.6**
DerivAC	75.1	78.3	**77.9**	**81.4**	70.2	68.1	72.6	70.3	**78.2**	**75.9**	**81.8**	**82.3**	**83.2**	83.5
NormAC	**76.5**	**78.5**	77.7	80.8	**71.4**	**70.6**	**74.1**	**71.3**	78.1	74.9	**81.8**	82.1	83.1	83.2







Score Gap between Members and Non-members