ETH zürich | SPY Lab

Blind Baselines Beat Membership Inference Attacks for Foundation Models

Debeshee Das, Jie Zhang, Florian Tramèr

1. Membership Inference Attack

2. MIA on Foundation Models

Constructing a post-hoc MIA Evaluation Dataset:

Vision Models

(with some debiasing post-processing)

Members: Dataset A

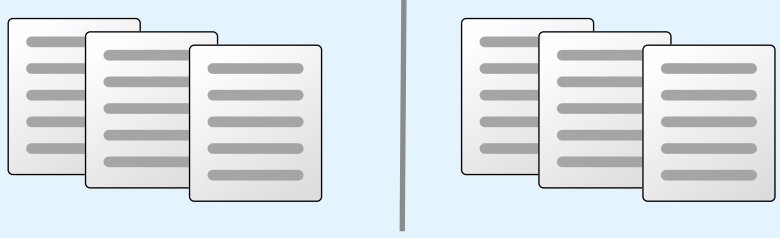
Non-Members: Dataset B

LLMs

Model Cutoff

Members: Before Cutoff

Non-Members: After Cutoff



Are these datasets really IID?

3. Blind Attack Baseline

Date Detection

Bag of Words Classifier

Greedy Rare Word Selection

4. Case Study Results

MI Dataset	Metric	Best Reported MIA(%)	Ours (%)
Dataset 1	TPR@5%FPR	43.2	**94.7**
Dataset 2	AUC ROC	88.0	**91.4**
Dataset 3	AUC ROC	79.6	**79.9**
Dataset 4	AUC ROC	74.5	**75.3**
Dataset 5	TPR@1%FPR	5.9	**10.6**
Dataset 6	TPR@1%FPR	2.5	**2.7**
Dataset 7	TPR@1%FPR	2.5	**8.9**
Dataset 8	TPR@1%FPR	18.8	**55.1**

5. Conclusion

1. Members and non-members of post-hoc MIA datasets can be reliably distinguished by simple blind attacks
2. Current evaluations of MI attacks for foundation models cannot be trusted
3. Datasets with IID train-test split like The Pile and DataComp should be used for MIA Evaluation