The Canary's Echo: Auditing Privacy Risks of LLM-Generated Synthetic Text

ICML 2025

Matthieu Meeus[1,2], Lukas Wutschitz[2], Santiago Zanella-Béguelin[2], Shruti Tople[2] and Reza Shokri[2,3]

(1) Imperial College London, (2) Microsoft, (3) National University of Singapore

Synthetic text data

Similar utility, yet not directly traceable to any original record

Original, sensitive data

Real text	Label
'very well-written and very well-acted.'	Positive
'two guys who desperately want to be quentin tarantino when they grow up'	Negative
'is a pan-american movie, with moments of genuine insight into the urban heart.'	Positive

Generate synthetic data

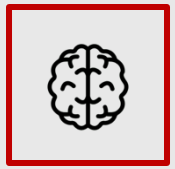
Synthetic data

Synthetic text	Label
'one of the most enjoyable romantic comedies of the year'	Positive
'if i 'm going to watch a three hour movie , i'd want it to be better than just good.'	Negative
'the only time when a remake of a classic is every bit as good as the original'	Positive

Use for downstream tasks

1

Finetune a pretrained LLM on the real data



2

Sample from the finetuned LLM

Auditing via Membership Inference Attacks (MIAs)

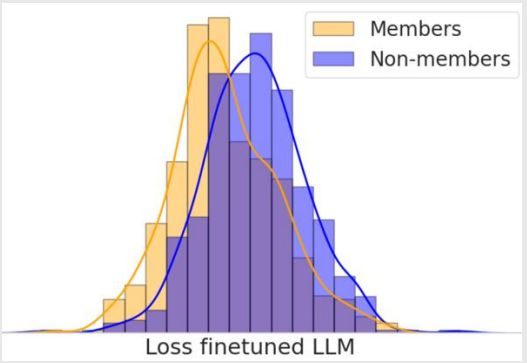
- While synthetic data is not directly traceable to original data, it does not mean it is free from any privacy risk.
- MIAs aim to infer if a given target sequence was part of the private dataset used to train a certain algorithm.

Canary

'When in comes times of turmoil... whats on sale and more important when, is best, this…'

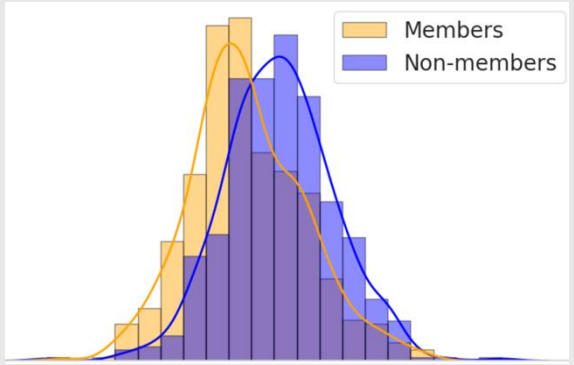
LLM

Attacker to predict member or non-member based on the model loss



Data-based MIAs

- We develop MIAs that do not rely on access to the model, but exclusively on the generated synthetic data.



?

Canary

'When in comes times of turmoil... whats on sale and more important when, is best, this…'

Synthetic data

Synthetic text	Label
'one of the most enjoyable romantic comedies of the year'	Positive
'if i 'm going to watch a three hour movie , i'd want it to be better than just good.'	Negative
'the only time when a remake of a classic is every bit as good as the original'	Positive

Attacker to predict member or non-member based on:

1

N-gram: compute the loss of the canary using an n-gram model trained on the synthetic data.

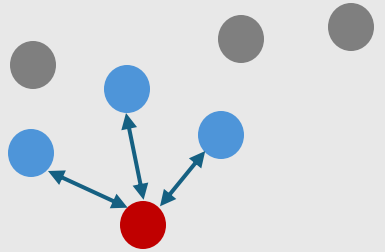
$$P(\text{times}|comes) = \frac{Count(comes\ times)}{Count(comes)}$$

Probability target sequence

Computed on synthetic

2

Similarity: compute the mean similarity between the canary to the closest synthetic sequences.



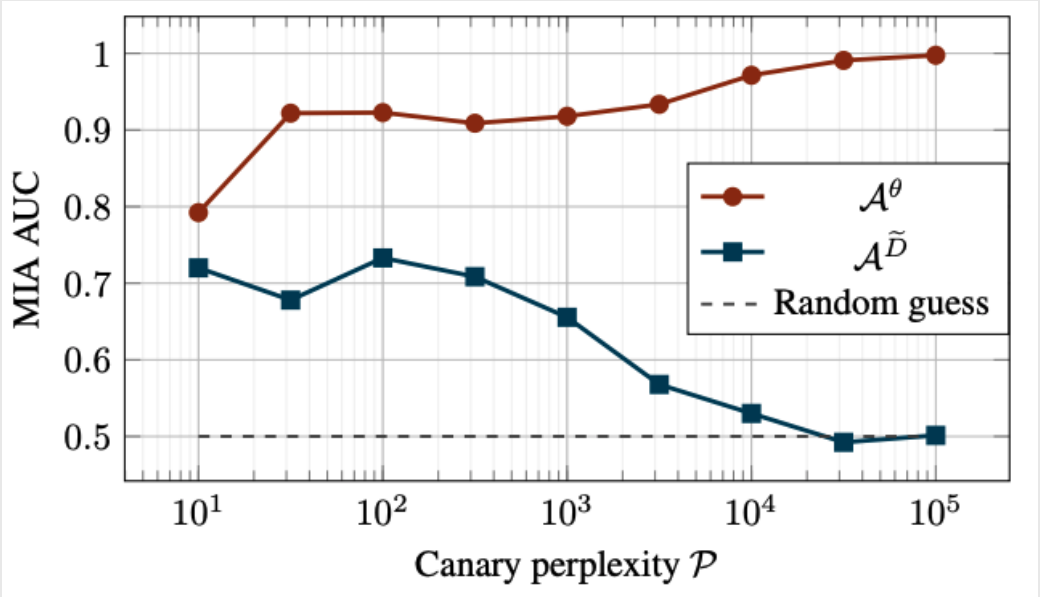
Synthetic data leaks private information

- **MIAs just based on synthetic data reach AUCs of 0.7+** (2-gram method works the best)
  - Synthetic data does leak private information!
  - To match the vulnerability of **model-based MIAs**, our data-only attacks need canaries to appear ~8× more often in the training data.

Dataset	Canary injection		ROC AUC			
	Source	Label	Model $\mathcal{A}^{\theta}$	Synthetic $\mathcal{A}^{\tilde{D}}$ (2-gram)	Synthetic $\mathcal{A}^{\tilde{D}}$ (SIM_{Jac})	Synthetic $\mathcal{A}^{\tilde{D}}$ (SIM_{emb})
SST-2	In-distribution[1]		0.911	0.741	0.602	0.586
	Synthetic	Natural	0.999	0.620	0.547	0.530
		Artificial	0.999	0.682	0.552	0.539
AG News	In-distribution		0.993	0.676	0.590	0.565
	Synthetic	Natural	0.996	0.654	0.552	0.506
		Artificial	0.999	0.672	0.560	0.525
SNLI	In-distribution[1]		0.892	0.718	0.644	0.630
	Synthetic	Natural	0.998	0.534	0.486	0.488
		Artificial	0.997	0.770	0.602	0.571

Traditional canaries fail for auditing synthetic data

- We vary the *perplexity* of the canaries we include;
- We find a novel trade-off
  - Model-based MIAs improve as canary perplexity increases.
  - Data-based MIAs work better for low perplexity, in-distribution canaries.



While rare canaries are memorized more by the model, their signal does not echo through the generated text

A new canary design to audit synthetic text

- We propose a new canary design with
  - an in-distribution prefix F (more easily echoed through the synthetic data)
  - high-perplexity suffix (better memorized by the model)
- Canaries with an in-distribution prefix *0 < F < max* work better for data-based MIAs!

Dataset	F	ROC AUC
SST-2	0	0.673
	10	0.715
	20	0.725
	30	**0.760**
	max	0.741
AG News	0	0.692
	10	0.646
	20	**0.716**
	30	0.710
	max	0.676

Conclusion

We propose an end-to-end pipeline to audit the privacy risks in LLM-generated synthetic text, with novel MIA techniques and optimal canaries.

We hope this enables making informed decisions about releasing synthetic data in practice.

More details in the paper!