Black-box Membership Inference Attacks on Synthetic Text via N-gram Overlap

Yidan Sun [1] **Viktor Schlegel** [1 2] **Srinivasan Nandakumar** [1] **Iqra Zahid** [1] **Siew-kei Lam** [3] **Anil A Bharath** [1]

[1] *Imperial College London, Imperial Global Singapore* [2] *University of Manchester, United Kingdom* [3] *Nanyang Technological University, Singapore*. Correspondence to: Yidan Sun ysun1@imperial.ac.uk.

1. Introduction

Synthetic text generation enables data sharing while aiming to prevent direct privacy leaks, often paired with Differential Privacy (DP) [1] for worst-case protection at the record level. It is widely used in healthcare [2], finance [3], and law [4], where real documents are restricted. Yet, recent evidence shows that even high-quality or nominally DP-protected synthetic data can leak information through subtle patterns or direct overlaps [5, 6], motivating closer empirical scrutiny.

Membership inference attacks (MIA). MIA test whether a given text was included in a model's training set or a generator's input; success undermines intended privacy guarantees. We study transparent, easily interpretable n-gram overlap as a minimal, model-agnostic attack on synthetic text releases.

Motivation and scope. MIA on synthetic data matters because synthetic corpora are often treated as de-identified, while text is sparse and uniquely phrased, making fine-grained overlaps informative. Attackers can operate in a realistic black-box setting with limited domain references, and practical DP deployments (e.g., large ε or partial coverage) may leave outputs empirically vulnerable. We ask: (i) can simple n-gram overlap suffice for membership inference? (ii) how does access to domain reference data affect attack strength? and (iii) how consistent are results across domains, generators, and nominal DP guarantees?

Impact. Even simple black-box attacks can pose risk to synthetic text, including under commonly used DP mechanisms. Organizations should treat DP-protected synthetic corpora as *auditable*, adopt routine pre-release checks, and consider mitigations (e.g., overlap filtering or stricter DP configurations) proportionate to measured risk.

2. Threat Model and N-gram MIA Approach

Consider a released synthetic text dataset S. The attacker's goal is, given a target document $x^\star$, to infer whether $x^\star$ was in the training data D used to produce S. We assume a black-box adversary with no access to the generator's internals and consider two access regimes: *(i) No reference:* the attacker observes only S; *(ii) Reference:* in addition to S, the attacker also has a real-world reference set R which is independent identical distribution with D, from the same domain.

N-gram MIA pipeline. We assume a set of target records $\{x_i\}$, where each x_i may or may not be a member of a private dataset D_{prv}. Both D_{prv} and an auxiliary reference dataset D_{ref} are obtained as i.i.d. samples from the same underlying corpus distribution, so that D_{ref} reflects typical domain statistics but does not contain any private samples.

We consider a black-box setting in which an analyst trains a synthetic text generator on D_{prv} and releases only the resulting synthetic dataset $\hat{D}_{\text{prv}}$. In our experiments, we instantiate this step with one representative state-of-the-art synthetic text generation methods, DP-gen [7], under various privacy budgets ϵ. **DP-gen** [7] trains Transformer language or seq2seq models with DP-SGD (per-example clipping and noise) and then generates synthetic text via standard decoding. The adversary observes $\hat{D}_{\text{prv}}$, the set of targets $\{x_i\}$, and optionally the reference dataset D_{ref}.

To mount an n-gram membership inference attack, we proceed as follows. For each released synthetic dataset $\hat{D}_{\text{prv}}$ we fit an n-gram language model and use it to estimate the likelihood $P_{\text{ng}}(x_i)$ for each target record x_i. When a reference dataset is available, we also train one or more n-gram models on D_{ref} and use them to compute a baseline likelihood $\bar{P}_{\text{ref}}(x_i)$ that captures how probable x_i is under typical domain statistics. Our membership score is then defined as a function of these quantities, for example

$$\begin{aligned} \Delta P(x_i) &= P_{\text{ng}}(x_i) - \bar{P}_{\text{ref}}(x_i) \quad \text{(with reference)}, \\ s_{\text{ng}}(x_i) &= P_{\text{ng}}(x_i) \quad \text{(without reference)}. \end{aligned} \tag{1}$$

Intuitively, if a target record x_i has been memorized or closely preserved by the synthetic generator, the n-gram model trained on $\hat{D}_{\text{prv}}$ will assign it an unusually high likelihood compared to what is expected from D_{ref}, leading to a large positive $\Delta P(x_i)$. We interpret larger scores as stronger evidence of membership and evaluate the resulting attack by computing ROC curves and summary metrics such as AUC over many trials in which membership of each x_i in D_{prv} is known.

With vs. Without Reference. If reference data R are available, we can calibrate what "typical" overlap looks like by comparing to overlaps between $x^\star$ and R, not just S.

3. Experiments

We evaluate n-gram MIA on multiple real-world datasets drawn from (1) PsyTAR:[8] (adverse drug effect detection in social media posts), (2)N2C2'08[9] (obesity and co-morbidity recognition in clinical dis-

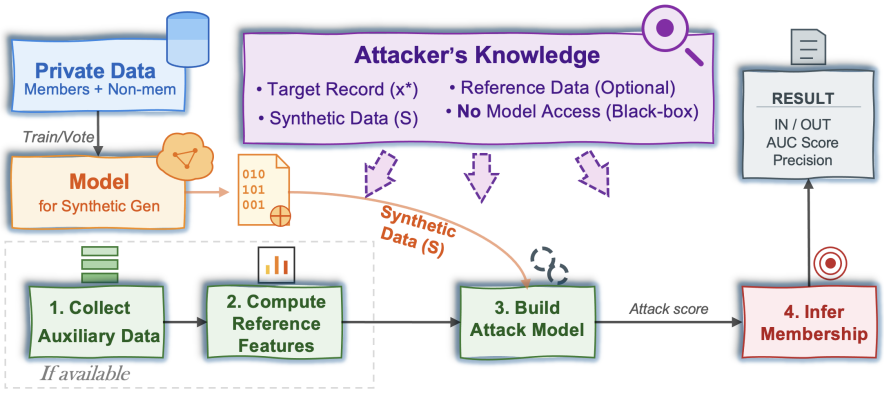

Fig. 1: Membership inference attack (MIA) pipeline: private data train a model or generator; an attacker inspects outputs or synthetic datasets (optionally with a domain reference set), computes features (e.g., n-gram overlap), and applies an attack model to decide membership and quantify risk.

Table 1: Membership Inference Attack performance (ROC-AUC, %, mean [95% CI]) on three datasets and DP synthetics from $\epsilon = \infty$ (non-DP) to 0.5.

Dataset	ϵ	ngram (ref)	ngram (no-ref)
PSYTAR ($\lvert D\rvert = 1000$)	∞	**0.65 [0.64, 0.66]**	0.62 [0.61, 0.63]
	4	0.39 [0.38, 0.40]	**0.57 [0.56, 0.58]**
	2	0.44 [0.43, 0.45]	**0.60 [0.59, 0.61]**
	1	0.57 [0.56, 0.58]	**0.59 [0.58, 0.60]**
	0.5	0.58 [0.57, 0.59]	**0.59 [0.58, 0.60]**
N2C2'08 ($\lvert D\rvert = 500$)	∞	**0.59 [0.59, 0.60]**	0.47 [0.46, 0.48]
	4	**0.48 [0.47, 0.49]**	0.45 [0.44, 0.47]
	2	**0.47 [0.46, 0.47]**	0.46 [0.45, 0.47]
	1	**0.46 [0.45, 0.47]**	0.45 [0.44, 0.46]
	0.5	**0.46 [0.45, 0.47]**	0.46 [0.44, 0.47]
DMSAFN ($\lvert D\rvert = 1000$)	∞	**0.82 [0.81, 0.82]**	0.56 [0.55, 0.57]
	4	**0.54 [0.53, 0.55]**	0.53 [0.52, 0.54]
	2	0.49 [0.48, 0.50]	**0.53 [0.52, 0.54]**
	1	**0.53 [0.52, 0.54]**	**0.53 [0.52, 0.54]**
	0.5	**0.56 [0.55, 0.57]**	0.54 [0.53, 0.55]

Each cell: ROC-AUC (%) with 95% confidence intervals over all audit trials (#trials = 100×(1+negatives per positive)), by bootstrap resampling.
ngram-ref: ngram similarity using reference set; **ngram:** ngram similarity, no reference.
Results shown for non-DP ($\epsilon = \infty$) and DP ($\epsilon = 4, 2, 1, 0.5$).

charge summaries) and (3) DMSAFN[1] (financial news). We measure ROC-AUC: the ability to separate members from non-members.

3.1 Main Results: Effectiveness of n-gram MIA

Table 1 shows that simple n-gram MIAs achieve consistently above-chance ROC-AUC across datasets and privacy levels. The attack is strongest on DMSAFN, reaching 0.82 AUC in the non-DP case and remaining around 0.53–0.56 even at $\epsilon = 0.5$, which indicates a clear and persistent membership signal in this financial domain. On PSYTAR and N2C2'08, AUCs are more moderate (typically 0.46–0.65), but still indicate non-trivial leakage under both non-DP and DP settings, suggesting that even noisy synthetic releases can preserve exploitable n-gram structure. Overall, the results highlight that high utility synthetic text is accompanied by measurable privacy risk, even when generation is performed under formal DP guarantees.

3.2 Discussion: With vs. Without Reference

The comparison of "ngram (ref)" and "ngram (no-ref)" reveals that reference data do not help uniformly. On PSYTAR and DMSAFN, reference-based attacks are clearly stronger in the non-DP regime, but the no-reference variant can be competitive or better at smaller ϵ. For N2C2'08, the reference-based attack is slightly better across all ϵ. These patterns suggest that reference calibration can significantly amplify risk in some domains, while synthetic-only attacks remain a realistic baseline threat even without auxiliary data.

4. Conclusion and Recommendations

N-gram overlap is a simple yet effective black-box MIA for synthetic text, particularly when the adversary has access to reference data. The resulting risk depends on both the underlying domain and the synthetic generation method, including the applied DP budget. Practical mitigations should account for surface-form overlap between released synthetic texts and plausible private inputs, rather than relying solely on the fact that generation occurs through a black-box model or nominal DP guarantees.

Acknowledgments

This research is part of the IN-CYPHER programme and is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. We are grateful for the support provided by Research IT in form of access to the Computational Shared Facility at The University of Manchester and the computational facilities at the Imperial College Research Computing

[1] https://huggingface.co/datasets/Daniel-ML/sentiment-analysis-for-financial-news-v2

Service[2].

References

[1] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[2] Alistair E.W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1), 12 2023.

[3] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

[4] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-scale multi-label text classification on eu legislation. *arXiv preprint arXiv:1906.02192*, 2019.

[5] Yidan Sun, Viktor Schlegel, Srinivasan Nandakumar, Iqra Zahid, Yuping Wu, Yulong Wu, Hao Li, Jie Zhang, Warren Del-Pinto, Goran Nenadic, et al. Synbench: A benchmark for differentially private text generation. *arXiv preprint arXiv:2509.14594*, 2025.

[6] Matthieu Meeus, Lukas Wutschitz, Santiago Zanella-Béguelin, Shruti Tople, and Reza Shokri. The canary's echo: Auditing privacy risks of llm-generated synthetic text. *arXiv preprint arXiv:2502.14921*, 2025.

[7] Lukas Wutschitz, Huseyin A. Inan, and Andre Manoel. dp-transformers: Training transformer models with differential privacy. `https://www.microsoft.com/en-us/research/project/dp-transformers`, August 2022.

[8] Maryam Zolnoori, Kin Wah Fung, Timothy B Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Yi Shuan Shirley Wu, Christina E Eldredge, Jake Luo, Mike Conway, et al. A systematic approach for developing a corpus of patient reported adverse drug events: a case study for ssri and snri medications. *Journal of biomedical informatics*, 90:103091, 2019.

[9] Ozlem Uzuner. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16(4):561–570, 07 2009.

[2]DOI: `https://doi.org/10.14469/hpc/2232`