

Who Bears Privacy Risk? Model-Dependent Memorization in Synthetic Tabular Data

Aidan LoStracco, Joshua August Skorburg, Wenjing Zhang

University of Guelph
lostraca@uoguelph.ca

Abstract

Privacy risk in synthetic tabular data is typically evaluated as an aggregate property: how much leakage exists, rather than who is exposed. But aggregate metrics treat individuals as undifferentiated, hiding whether certain records or subgroups are systematically more exposed than others. If risk is unevenly distributed, a natural but largely unexplored question is whether the choice of generative model itself shifts who bears it. We address this question through per-sample nearest-neighbor memorization analysis across diffusion-based and LLM-based tabular generators on real-world financial and clinical datasets. We find that the training records at highest memorization risk vary substantially across generators, with some model pairs exposing nearly disjoint victim sets, and that the most exposed demographic subgroups shift with the generator deployed: on the financial dataset, diffusion models disproportionately expose lower-income individuals while an LLM-based generator over-exposes higher-income ones; on the clinical dataset, LLM-based models concentrate memorization on a clinically sensitive subpopulation that diffusion models largely spare. Standard aggregate privacy metrics surface neither pattern; in at least one case, a model with more favorable aggregate metrics shifts who is exposed rather than uniformly reducing risk. We argue that model selection for synthetic data generation is implicitly a decision about which subpopulations bear privacy risk, constituting a fairness consideration distinct from aggregate privacy guarantees. Privacy evaluations that rely on aggregate metrics alone may fail to identify the individuals most vulnerable to memorization-based leakage.

1 Introduction

Data in sensitive domains such as healthcare and finance is subject to strict privacy regulations, including HIPAA and GDPR, which significantly restrict how it can be shared and used. These restrictions, while necessary, limit research progress, hinder collaboration across institutions, and create barriers to AI model development. Synthetic data has emerged as a popular proposed solution to this tension (Jordan et al. 2022). Rather than sharing real records, organizations can instead share generated data that mirrors the statistical properties of the original dataset without exposing personally identifiable information. Recent advances in gener-

ative AI have made synthetic data increasingly realistic and useful, accelerating its adoption for research, model training, and benchmarking (Giuffrè and Shung 2023).

The core assumption underlying this approach is that synthetic data does not correspond to real individuals, and therefore can be treated as inherently safe or anonymized. This assumption is frequently used to justify bypassing stricter privacy controls; in research, industry, and regulatory contexts alike. In this paper, we challenge that assumption. While prior work has characterized memorization within individual model classes (Fang et al. 2025a,b; Meeus et al. 2024), no existing study examines whether different models expose different individuals, or what this implies for who bears privacy risk when practitioners choose between them. We address this gap through an empirical comparison of memorization in tabular diffusion models and transformer-based generators across two real-world datasets. We find that model selection is not privacy-neutral: different generators expose largely disjoint sets of individuals, the demographic subgroups most at risk shift substantially depending on which model is deployed, and standard aggregate privacy metrics fail to surface either pattern. We argue that privacy evaluation for synthetic data should be understood as a distributional fairness problem, not only a security one. This reframing is directly relevant to trustworthy AI, which requires not just accurate or useful systems but systems whose risks and burdens are distributed equitably across the populations they affect.

To the best of our knowledge, this is the first work to empirically compare per-sample memorization across architectural classes of tabular generative models and to characterize the demographic composition of the resulting victim sets. Our contributions are threefold:

1. We show that standard aggregate privacy metrics produce inconsistent model rankings and fail to surface which individuals are exposed, motivating per-sample analysis as a necessary complement to existing audit practice.
2. We provide the first empirical comparison of per-sample memorization across diffusion-based and LLM-based tabular generators, showing that victim sets vary substantially across generators and that this variation is not reliably predicted by architectural class, a pattern prior work evaluating memorization within single model classes was not designed to detect.

3. We characterize the demographic composition of victim sets across both model classes and two datasets, showing that the subpopulations bearing the highest memorization risk shift substantially depending on which model is deployed.

2 Background and Related Work

The background is organized to build the case that who bears memorization risk is a meaningful and currently invisible question. Section 2.1 describes the two model classes we evaluate and the properties relevant to how each learns from training data. Section 2.2 reviews the memorization literature, establishing that per-sample vulnerability is a meaningful and measurable phenomenon rather than a theoretical concern. Section 2.3 examines the architectural mechanisms by which diffusion and LLM-based models differ in how they fit training records, motivating the possibility that different generators expose different individuals. Section 2.4 characterizes current evaluation practice and its aggregate-metric limitations, establishing that existing audits are not designed to surface this. Section 2.5 frames the distributional and fairness dimension of that gap: the question of who bears memorization risk is distinct from how much risk exists in aggregate, and the existing literature has not addressed it across model classes.

2.1 Synthetic Data Generation

Synthetic data models learn underlying data distributions rather than storing or reproducing individual records. This is essential for maintaining utility because if the generated data does not realistically reflect the original, it is not useful. For example, a model trained on electronic health records may learn relationships between attributes such as age, diagnosis, and treatment outcomes, enabling it to generate new patient records that appear realistic without explicitly reproducing any single individual.

Several generative modeling approaches have been proposed for synthetic data generation. Generative adversarial networks (GANs) were the dominant method for many years, particularly in image domains, but are notoriously difficult to train and evaluate, and have largely been superseded by more capable architectures (Wiatrak, Albrecht, and Nystrom 2020; Dhariwal and Nichol 2021). This paper focuses on two classes that represent the current state of the art for tabular data in sensitive domains: diffusion models and transformer-based language models adapted for tabular data.

Diffusion models generate data through an iterative denoising process. During training, Gaussian noise is progressively added to real data samples until the original signal is destroyed. The model learns to reverse this process, recovering realistic samples from noise. At generation time, the model starts from pure noise and applies learned denoising steps to produce a synthetic record. This approach has been extended to structured tabular data, where it operates over mixed feature spaces containing both continuous and categorical variables (Ho, Jain, and Abbeel 2020; Kotelnikov et al. 2023; Zhang et al. 2024).

Transformer-based language models adapted for tabular data take a fundamentally different approach. Rather than

operating in feature space, these models serialize tabular records as natural language sequences, treating each row as a structured text prompt. Generation proceeds autoregressively: the model samples one token at a time, conditioned on all previously generated tokens. This framing allows general-purpose language model architectures to be applied directly to tabular data without domain-specific modifications (Vaswani et al. 2017; Borisov et al. 2023; Zhao, Birke, and Chen 2025).

Despite their architectural differences, both model classes are trained on real-world data distributions and are therefore not inherently insulated from privacy leakage. The nature of that leakage, however, differs in ways that are directly tied to how each model represents and generates data, a distinction we examine in Section 2.3.

2.2 Memorization

Memorization occurs when a generative model’s outputs remain closer to specific training records than the underlying distribution warrants, reflecting overfitting to individual examples rather than generalization from them. It is a well-documented phenomenon across generative model types: large language models have been shown to reproduce training data verbatim under certain prompting conditions (Carlini et al. 2021), membership inference attacks have demonstrated information leakage across a range of generative architectures (Shokri et al. 2017; Hu et al. 2022), and image diffusion models have been shown to memorize and reproduce training images with high fidelity (Somepalli et al. 2023; Carlini et al. 2023).

Until recently, comparatively little work had examined memorization in tabular generative models specifically. The first comprehensive study of this phenomenon in tabular settings was conducted by Fang et al., who introduced a per-sample memorization scoring framework based on nearest-neighbor distances in feature space (Fang et al. 2025a). Their framework identifies individual training records that are disproportionately close to generated samples, enabling memorization to be measured at the level of specific records rather than as a global dataset property. A subsequent data-centric analysis extends this framework, finding that memorization is not uniformly distributed: some records are highly exposed while the majority are not, a pattern with direct implications for how privacy risk should be evaluated and reported (Fang et al. 2025b).

A related thread of work goes further, asking not only whether memorization occurs but which records are most vulnerable. Meeus et al. propose a principled nearest-neighbor-based technique for identifying vulnerable records prior to synthetic data release, showing that targeted record-level analysis substantially outperforms aggregate evaluation methods (Meeus et al. 2024). Ward et al. find that data-copying in tabular generative models poses both privacy and fairness threats, with disproportionate copying concentrated among outlier and certain privileged subpopulations (Ward, Wang, and Cheng 2024). These results establish that the distribution of memorization risk across training records is a meaningful object of study and that aggregate evaluation metrics systematically obscure it.

Critically, memorization does not require exact reproduction of a training record to constitute a privacy violation. Partial reconstruction alone can expose sensitive information (Stadler, Oprisanu, and Troncoso 2022). For instance, a synthetic record that matches most attributes of a real individual while differing only slightly may still effectively reveal that person’s medical history or financial profile. Beyond direct similarity, memorization can also enable attribute inference attacks, where an adversary uses synthetic outputs to infer hidden attributes from partially observed data (Annamalai, Gadotti, and Rocher 2024). If a model has learned strong correlations between features, known information about an individual could be used to infer sensitive unknowns, such as inferring a specific treatment outcome from age and diagnosis. Together, these attack vectors indicate that privacy risk in synthetic data extends well beyond exact duplication.

What the existing literature leaves open is whether the distribution of memorization risk across training records shifts depending on which generative model is deployed, and whether the demographic subpopulations bearing the greatest risk differ as a consequence of that architectural choice. Fang et al. and Meeus et al. characterize per-record vulnerability within single model classes without asking whether different models expose different individuals (Fang et al. 2025a,b; Meeus et al. 2024). Ward et al. observe fairness implications of data-copying but do not compare which subpopulations are at risk across model classes (Ward, Wang, and Cheng 2024). This paper addresses that gap directly, using a common nearest-neighbor evaluation framework to compare per-sample memorization across diffusion and LLM-based tabular generators, with a focus on whether architecture choice determines who bears privacy risk.

2.3 Architecture-Dependent Memorization Mechanisms

The mechanisms by which generative models memorize training data are not uniform across architectures, and these differences have direct implications for which individuals are most at risk. Diffusion models and transformer-based language models differ fundamentally in how they represent and generate data, and understanding this distinction is necessary for reasoning about whose records are likely to be exposed under each approach.

Because diffusion models operate entirely in feature space, memorization reflects how faithfully the model replicates the training distribution. The most intensely memorized records tend to be those with unusual feature combinations that the model fits more tightly than the broader distribution can account for (Fang et al. 2025b). Nearest-neighbor distance metrics are a natural evaluation tool for this setting because both the generation process and its failure modes are defined in feature space (Stadler, Oprisanu, and Troncoso 2022; Fang et al. 2025a).

Because LLM-based generators serialize records as token sequences and generate autoregressively, a record’s vulnerability is not simply a function of where it sits in feature space. It is also a function of how its serialized representation behaves under autoregressive generation. Records with complete, consistently formatted feature vectors may

be more tightly fit because they provide cleaner training signal; records with unusual but coherent token sequences may be reproduced more faithfully than feature-space distance alone would predict. Ward et al. demonstrate that this sequential generation process produces memorization failure modes qualitatively distinct from those of diffusion models (Ward et al. 2025). Importantly, models within the same architectural class are not necessarily homogeneous in their memorization behavior, and the degree to which any two models expose similar individuals may depend on the dataset as much as on their shared architecture.

This architectural divergence raises a concrete empirical question, though one without an obvious answer. Diffusion models and LLM-based generators fit training records through different mechanisms, but it is not clear a priori that this produces different victim sets; the subset of training records most exposed by a given model. Both classes might converge on the same individuals regardless of mechanism, since records that are outliers in feature space may simultaneously be unusual in their serialized representations. Whether architectural differences produce meaningful divergence in victim sets, and whether any such divergence falls along demographic lines, is therefore an empirical question rather than a predictable consequence of architectural difference. This paper investigates it directly, applying a common nearest-neighbor evaluation framework to both model classes and comparing the resulting victim sets and their demographic composition across two datasets.

2.4 Current Privacy Evaluation Metrics

Determining whether a synthetic dataset is sufficiently safe requires evaluation frameworks capable of detecting the privacy risks described above. Several approaches have been developed to assess this, each making different assumptions about what constitutes a privacy violation and how it can be measured.

The most common methods are membership inference attacks, similarity metrics, and statistical comparisons. Membership inference attacks test whether a given record can be identified as a member of the training set (Shokri et al. 2017). Similarity metrics measure the distance between synthetic and real records. For instance, a nearest neighbor metric can flag synthetic samples that are significantly closer to their nearest neighbor than their second nearest neighbor in the feature space, suggesting likely memorization (Fang et al. 2025a). Statistical comparisons measure distributional similarity between the synthetic and real datasets, primarily to ensure utility rather than privacy, though they are sometimes conflated with privacy evaluation.

Differential privacy (DP) offers a more formal approach, providing mathematical guarantees that limit the influence of any single training record on the model’s output (Dwork 2008; Dwork and Roth 2014). The strength of these guarantees is controlled by a privacy budget ϵ , where smaller values provide stronger protection at the cost of greater distortion to the model’s outputs. This tradeoff between privacy and utility is unavoidable in practice: ϵ is often set permissively to preserve data utility, substantially weakening the per-record protection DP provides in theory.

These evaluation practices are foundational to how synthetic data is trusted in practice. They are used in academic research to justify safety claims, in industry tools to make privacy guarantees to customers, and in regulatory contexts to justify bypassing data consent requirements. However, they share a common limitation: they predominantly report aggregate metrics, such as average similarity scores or overall attack success rates. While this reflects practical constraints, it means evaluation focuses on average-case performance rather than worst-case exposure.

Some tail exposure is an inherent consequence of the privacy-utility tradeoff that all approaches face, including DP. A model that memorizes no training records at all would produce synthetic data with no statistical fidelity to the original distribution. In practice, any generator that preserves enough utility to be useful will expose some subset of individuals to elevated memorization risk. The question is therefore not whether a 1% tail exists, but who occupies it. Even if a generated dataset contains 99% well-protected records, the 1% that are exposed still represent real individuals whose sensitive information has been compromised, and for those individuals, the consequences are no less severe for being statistically rare. Aggregate metrics do not surface this: they cannot reveal whether that exposed minority is drawn uniformly from the population or concentrated among individuals who are already vulnerable by virtue of their clinical profiles, financial circumstances, or demographic characteristics. Nor do they reveal whether that concentration shifts when a different model is deployed. These are the questions this paper addresses.

2.5 Distributional Risk and Who Bears It

Privacy risk in synthetic data is not uniformly distributed across all records. Recent work has shown that memorization often follows a heavy-tailed pattern, where a small subset of training examples accounts for a disproportionately large share of overall privacy risk (Fang et al. 2025b). While many individuals in a dataset may be well-protected, a minority of records, particularly those that are rare or distinctive, experience significantly higher exposure.

In practice, these high-risk individuals often correspond to outliers in the data distribution. In healthcare settings, this includes patients with rare diseases, uncommon combinations of clinical attributes, or atypical monitoring histories that set them apart from the majority of the training population. In financial settings, analogous cases arise for individuals with unusual credit histories, atypical payment patterns, or demographic characteristics that place them at the margins of the training distribution. In both domains, the individuals most at risk are precisely those whose records are hardest to generalize from, and whose exposure carries the greatest potential for harm.

This uneven distribution of privacy risk is directly analogous to well-documented disparities in model performance across demographic groups in machine learning (Buolamwini and Gebru 2018; Obermeyer et al. 2019). Rather than differences in predictive accuracy, however, the disparity here is in privacy protection. Individuals who are already atypical or underrepresented face disproportionately greater

exposure, compounding existing vulnerabilities rather than distributing risk equitably. This reframing has an important implication: privacy in synthetic data should be understood as a distributional and fairness-related issue, not a purely global property of the dataset.

The architecture-dependent mechanisms described in Section 2.3 add a further dimension to this distributional picture. If diffusion models and LLM-based generators expose individuals through different memorization mechanisms, the subpopulations bearing the greatest risk may not be the same under each architecture, and the degree to which they differ may itself depend on the dataset and the specific models involved. A record that is an outlier in feature space may be highly exposed by one model while remaining largely protected by another; a record with an unusual but complete feature profile may be tightly memorized by an autoregressive generator precisely because of the properties that make it clinically or financially distinctive. Whether these architectural differences produce systematic demographic divergence in victim sets is the empirical question this paper investigates. If they do, then the choice of generative architecture is not a purely technical decision: it is a decision with distributional consequences, determining in part who bears privacy risk and whether that risk is visible to those responsible for managing it. Practitioners and institutions deploying synthetic data pipelines without accounting for these differences may inadvertently concentrate privacy burdens on the most vulnerable members of a dataset, without any awareness that this has occurred.

3 Methodology

Datasets. We evaluate on two datasets from sensitive domains with distinct feature structures and demographic compositions. The UCI Default of Credit Card Clients (UCI) dataset contains 30,000 records, split 80/10/10 into training (24,000), validation, and test sets. It includes 23 features covering credit limit, payment history, and demographic attributes including sex, education, and marital status, with a binary default outcome (22.2% positive rate). The Diabetes 130-US dataset contains 101,766 clinical encounters from 130 US hospitals. We remove identifier columns to prevent trivial memorization of non-informative features, features with extreme missingness where imputation would introduce artifacts, and near-constant medication variables that contribute negligible discriminative signal, yielding 35 informative features after preprocessing. Because multiple encounters for the same patient would inflate memorization scores by placing copies of a patient’s record in both training and test sets, we deduplicate to each patient’s most recent encounter, yielding 69,985 unique records split 80/10/10 into training (55,989), validation, and test sets. Features include ICD-9 diagnosis codes, medication indicators, and demographic attributes including race, age, and HbA1c measurement status, with a binary 30-day readmission outcome. The two datasets were chosen to test whether architecture-dependent memorization patterns generalize across domains: UCI provides a financial context with income and education as primary demographic axes, while Diabetes 130-US provides a clinical context where

privacy stakes are higher and demographic axes include race and clinical subgroup membership.

Models. We evaluate four tabular generative models representing two architectural classes. In the diffusion class, TabDDPM uses an MLP-based denoising network with Gaussian diffusion for continuous features and multinomial diffusion for categorical features (Kotelnikov et al. 2023). TabSyn uses a two-stage approach: a VAE encodes mixed-type tabular records into a continuous latent space, which a transformer-based diffusion model then learns to generate (Zhang et al. 2024). In the LLM-based class, GReaT serializes tabular records as natural language sequences and fine-tunes a pretrained GPT-2 model using autoregressive next-token prediction (Borisov et al. 2023). Tabula uses the same autoregressive text serialization paradigm but trains an equivalent GPT-2 architecture from randomly initialized weights rather than fine-tuning a pretrained model (Zhao, Birke, and Chen 2025). These four models were selected as representative, widely used instantiations of each architectural class, with strong reported performance on tabular data benchmarks. The inclusion of both GReaT and Tabula helps distinguish effects associated with pretrained language-model initialization from effects associated with autoregressive text serialization. All models are trained on three random seeds per dataset; memorization scores are computed independently per seed and aggregated. For each seed, we generate a synthetic dataset matching the size of the corresponding training split. Synthetic sample counts were therefore fixed across models and seeds. TabDDPM is excluded from Diabetes 130-US results due to an architectural failure described in Section 4.2. All models are trained using the hyperparameter settings and preprocessing configurations recommended in their original implementations. We did not perform architecture-specific privacy tuning or differential privacy training for any model. All experiments were conducted in Python on a system with an Intel Core Ultra 7 265F CPU, 64GB RAM, an NVIDIA RTX 5090 GPU, and 2TB SSD storage.

Aggregate Privacy Metrics. As a baseline for comparison, we compute three standard aggregate privacy metrics representing distinct audit perspectives on memorization risk. The Distance to Closest Record (DCR) measures output-side risk: for each synthetic record, its minimum Gower distance to any training record, reflecting how closely the synthetic dataset approaches real individuals on average. The Nearest Neighbour Distance Ratio (NNDR) measures concentration of that proximity: for each synthetic record, the ratio of its distance to the nearest training record versus the second-nearest, distinguishing sharp memorization of specific individuals from diffuse distributional replication. We also report a memorization exposure rate derived from the per-sample scores described below: for a given threshold τ , the true positive rate is the fraction of training records with $\text{mem_score} > \tau$, quantifying attacker advantage from the input side. DCR and NNDR are reported as means across all synthetic records, averaged over three generation seeds. These metrics are included not as the primary evaluation framework but as a point of contrast: we show

that a practitioner running all three standard audit perspectives can still reach incorrect conclusions about which model poses greater privacy risk without taking into account who is being exposed, motivating the per-sample analysis that follows.

Memorization Scoring. We use the per-sample memorization scoring framework introduced by (Fang et al. 2025a), adapted for mixed-type tabular data using Gower distance. For each training record x_i , we compute:

$$\text{syn_score}(x_i) = \frac{1}{k} \sum_{j=1}^k d_G(x_i, \hat{x}_j^{(i)}) \quad (1)$$

where $\hat{x}_1^{(i)}, \dots, \hat{x}_k^{(i)}$ are the k nearest synthetic records to x_i under Gower distance d_G and:

$$\text{ref_score}(x_i) = d_G(x_i, x_{(1)}^{\text{test}}) \quad (2)$$

where $x_{(1)}^{\text{test}}$ is the nearest test-set record to x_i . The memorization score is then:

$$\text{mem_score}(x_i) = \text{ref_score}(x_i) - \text{syn_score}(x_i) \quad (3)$$

A positive mem_score indicates that the synthetic distribution covers x_i more closely than the held-out test set does, suggesting the record was memorized rather than generalized from. Crucially, ref_score depends only on the train/test split and is identical across models, meaning mem_score controls for a record’s intrinsic isolation in feature space rather than simply flagging outliers. Gower distance is used throughout because it handles mixed continuous and categorical features without requiring manual encoding decisions, normalizing each feature to $[0, 1]$ before computing a weighted mean absolute difference. Scores are computed across three generation seeds per model and averaged. Results are stable across $k \in \{1, 3, 5, 10, 20\}$; we report $k = 5$ throughout.

Victim Sets and Cross-Model Overlap. For each model, we define victim sets as the top $p\%$ of training records by mem_score , representing the records at greatest memorization risk. We compute victim sets at $p \in \{1\%, 5\%, 10\%\}$; the within-class versus cross-class overlap pattern emerges at $p = 1\%$ and strengthens monotonically, and is absent below $p = 0.5\%$ where victim sets are too small for stable estimates. To quantify whether different model classes expose different individuals, we compute pairwise Jaccard similarity between victim sets:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

We compare within-class pairs (diffusion-diffusion, LLM-LLM) against cross-class pairs (diffusion-LLM) and report lift over a random baseline, where the null Jaccard for two sets of size n drawn uniformly from N records is approximately $n/(2N - n)$. High within-class lift relative to cross-class lift indicates that which individuals are exposed depends substantially on which model class is deployed. We

also characterize exclusive victim sets, records appearing in the top $p\%$ of one model class but not the other, and analyze their demographic composition directly.

Demographic Analysis. For each model and victim set threshold, we compare the demographic composition of victim records against the full training set base rate. For each demographic category c , we compute an enrichment ratio:

$$\text{enrichment}(c) = \frac{\text{victim rate}(c)}{\text{base rate}(c)} \quad (5)$$

where values above 1 indicate overrepresentation among memorization victims relative to the training population. We test for statistically significant demographic skew using chi-square goodness-of-fit tests per feature per model-threshold combination, and Mann-Whitney U tests to compare `mem_score` distributions across demographic groups. Demographic dimensions analyzed on UCI include sex, education, marital status, age group, credit limit quartile, and default status. On Diabetes 130-US they include race, gender, age group, HbA1c measurement status, insulin use, and readmission status.

4 Results

We report results across four subsections, structured to build the central argument in sequence. Section 4.1 establishes that standard aggregate metrics produce inconsistent model rankings and fail to surface subgroup-level patterns, motivating why per-sample analysis is necessary rather than supplementary. Section 4.2 characterizes per-sample memorization score distributions, confirming that the scoring framework produces meaningful, non-random signal on both datasets before any cross-model comparison is made. Section 4.3 presents the core empirical claim: that different models expose largely different individuals, with victim sets varying substantially across generators even when aggregate metrics suggest comparable risk. Section 4.4 examines the demographic composition of those victim sets, showing that the divergence documented in Section 4.3 falls along clinically and socioeconomically meaningful lines and that current evaluation practice gives practitioners no visibility into which subpopulations bear the burden of that exposure.

4.1 Aggregate Metrics Produce Inconsistent Risk Rankings

We first evaluate all four models using three standard aggregate audit perspectives: DCR, which measures how closely synthetic records approach training data on average; NNDR, which measures the sharpness of that proximity; and memorization exposure rate, which measures what fraction of training records are flagged as exposed at a given `mem_score` threshold. Table 1 reports results for both datasets.

The metrics disagree. On UCI, GReaT has the lowest DCR mean (0.0080), suggesting the highest risk by that measure, yet the lowest exposure rate at $\tau = 0.05$ (1.0%), suggesting the lowest risk by that measure. The threshold $\tau = 0.05$ was selected to reflect a meaningful margin above zero memorization signal. An auditor on UCI selecting the model with lowest exposure rate chooses GReaT.

That model has the strongest demographic disparity in our analysis: it under-exposes $\leq \$50k$ credit-limit individuals by 54% while over-exposing the \$140-240k bracket. The metric that recommended GReaT as safest selected the model exhibiting the strongest subgroup skew in our analysis. On Diabetes 130-US, Tabula leads on DCR while GReaT leads on exposure rate at $\tau = 0.05$. No metric produces a consistent ordering across both datasets or both perspectives. More critically, none of these values surface which individuals or subpopulations are being exposed: a limitation we address directly in the following sections.

4.2 Per-Sample Memorization Score Distributions

On UCI, all four models produce positive mean `mem_scores`, confirming that each generator produces synthetic records systematically closer to training data than the held-out test set. Diffusion models show the broadest memorization footprint: TabDDPM and TabSyn each exhibit positive `mem_score` for 88-89% of training records. GReaT covers substantially fewer records (70.2%) but its maximum `mem_score` (0.095) is only modestly below TabDDPM’s (0.154), indicating that the records GReaT does memorize are fit with comparable intensity. Tabula is intermediate in both coverage (85.8%) and mean score. A practitioner comparing these models on coverage alone would select GReaT as least risky, the same model that, as Section 4.4 shows, produces the most concentrated demographic skew of any model in our analysis.

On Diabetes 130-US, TabDDPM is excluded from results: its inability to handle high-cardinality ICD-9 diagnosis columns (up to 757 unique values per column) caused distribution failure, producing synthetic records so far from the training distribution that all `mem_scores` are negative and victim sets reflect encoding artifacts rather than genuine memorization. Among the three remaining models, overall memorization intensity is lower than on UCI, reflecting the larger training set (55,989 vs. 24,000 records) and higher-dimensional feature space. All three produce negative mean `mem_scores`; victim sets at each threshold therefore capture the least-negative tail rather than strongly positive records. TabSyn retains the broadest footprint (28.0% positive `mem_score`), while GReaT again exhibits the spikiest profile: lowest coverage (14.2%) combined with the highest maximum `mem_score` (0.110) of any valid model. The within-model pattern is consistent across both datasets: diffusion models memorize broadly, LLM-based models memorize fewer records but with concentrated intensity on those they do expose.

4.3 Cross-Model Victim Overlap

Table 2 reports pairwise Jaccard similarity between victim sets at $p = 1\%$ and $p = 5\%$ thresholds, with lift over a random null baseline. Every model pair exceeds the null by a substantial margin at both thresholds, confirming that memorization structure is real and non-random across all models.

The within-class versus cross-class pattern differs between datasets but holds in both. On UCI, the diffusion-diffusion pair (TabDDPM $\leftrightarrow$ TabSyn, $J = 0.326$, $63\times$

Metric	UCI Default Credit				Diabetes 130-US		
	TabDDPM	TabSyn	GReaT	Tabula	TabSyn	GReaT	Tabula
DCR mean	0.0213	0.0152	0.0080	0.0137	0.1457	0.1479	0.1341
NNDR mean	0.777	0.776	0.768	0.715	0.954	0.927	0.924
Exposure rate ($\tau=0.05$)	3.6%	3.3%	1.0%	2.4%	0.1%	0.5%	0.2%

Table 1: Aggregate privacy metrics across both datasets. No metric produces a consistent risk ordering across models. TabDDPM excluded from Diabetes 130-US (architectural failure; see Section 4.2).

Model A	Model B	UCI Default Credit				Diabetes 130-US			
		$p = 1\%$		$p = 5\%$		$p = 1\%$		$p = 5\%$	
		Jacc.	Lift	Jacc.	Lift	Jacc.	Lift	Jacc.	Lift
TabDDPM	TabSyn	0.326	63.4×	0.494	19.1×	—	—	—	—
GReaT	Tabula	0.209	40.6×	0.269	10.4×	0.181	35.5×	0.268	10.5×
TabSyn	GReaT	0.174	33.8×	0.227	8.8×	0.019	3.7×	0.081	3.2×
TabSyn	Tabula	0.294	57.1×	0.383	14.8×	0.016	3.2×	0.059	2.3×
TabDDPM	GReaT	0.148	28.8×	0.224	8.7×	—	—	—	—
TabDDPM	Tabula	0.311	60.6×	0.388	15.0×	—	—	—	—
Null (random)		0.005		0.026		0.005		0.026	

Table 2: Pairwise Jaccard similarity between victim sets at two percentile thresholds, with lift over random null. Within-class pairs in **bold**. TabDDPM excluded from Diabetes 130-US; see Section 4.2

lift at $p = 1\%$) shows the strongest overlap of any pair. GReaT $\leftrightarrow$ Tabula ($J = 0.209$, $41\times$) is the next highest but substantially lower. GReaT’s overlap with diffusion models ($J = 0.148$ - 0.174) is roughly half its within-class overlap, while Tabula’s overlap with diffusion models ($J = 0.294$ - 0.311) is comparable to diffusion-diffusion, placing Tabula closer to the diffusion class than to GReaT on this dataset. A practitioner who deployed Tabula expecting LLM-like victim sets on this dataset would instead be exposing a population closer in composition to what a diffusion model exposes.

On Diabetes 130-US the pattern reverses. GReaT $\leftrightarrow$ Tabula dominates ($J = 0.181$, $36\times$ lift), while cross-class pairs are near-null: TabSyn $\leftrightarrow$ GReaT ($J = 0.019$, $3.7\times$) and TabSyn $\leftrightarrow$ Tabula ($J = 0.016$, $3.2\times$) are an order of magnitude lower than the within-class LLM pair. Switching from TabSyn to GReaT on this dataset does not meaningfully shift memorization intensity, but it almost entirely changes which patients are exposed. Tabula here aligns with GReaT rather than with the diffusion model. The cross-class gap is more pronounced on Diabetes 130-US than on UCI, with LLM–LLM Jaccard approximately $10\times$ higher than either diffusion-LLM pair.

Across both datasets, which individuals face the highest memorization risk depends substantially on which model is deployed, a pattern that aggregate metrics do not surface and whose demographic consequences we examine in Section 4.4.

4.4 Demographic Composition of Victim Sets

Table 3 reports the demographic composition of $p = 5\%$ victim sets against training-set base rates for both datasets.

UCI: target variable exposure is universal; credit-limit enrichment inverts by model class. Across all four models, individuals who defaulted are substantially overrepresented in victim sets. At $p = 5\%$, default victim rates range from 38.9% (GReaT) to 42.8% (Tabula) against a base rate of 22.2%, with all chi-square $p < 10^{-19}$. Regardless of architecture, the minority class with the most distinctive feature patterns bears the highest memorization risk.

Credit-limit enrichment diverges sharply between model classes. GReaT under-exposes individuals with credit limits $\leq \$50k$ (11.8% victim rate vs. 25.7% base, -54%) while over-exposing the \$140-240k bracket (34.1% vs. 25.4% base). TabDDPM, TabSyn, and Tabula show the opposite: $\leq \$50k$ is over-exposed in all three (30.8-34.8%). One possibility is that GReaT over-generates high credit limits, which would leave low-limit training records without nearby synthetic neighbors and therefore unexposed. Confirming this mechanism is left to future work. Education follows the same axis: High School graduates are over-represented in diffusion and Tabula victim sets (23.2-29.3%) while GReaT over-exposes Graduate-educated individuals (46.4% vs. 35.3% base).

The practical consequence is visible in the exclusive victim sets. At $p = 5\%$, 611 records are memorized only by diffusion models and 896 only by LLM-based models. Diffusion-exclusive victims are lower-education and lower-income; LLM-exclusive victims are higher-education and mid-to-high income. Both groups are disproportionately

Dataset	Category	Base	Victim rate (base rate)			
			TabDDPM	TabSyn	GReaT	Tabula
UCI	Default (yes)	22.2%	42.1%	39.1%	38.9%	42.8%
	Education: HS	16.5%	29.3%	26.3%	15.9%	23.2%
	Education: Graduate	35.3%	30.9%	33.4%	46.4%	33.9%
	Credit $\leq$ \$50k	25.7%	34.8%	31.3%	11.8%	30.8%
	Credit \$140–240k	25.4%	21.3%	23.8%	34.1%	26.2%
D130	HbA1c: not tested	81.7%	—	82.5%	33.5%	30.9%
	HbA1c: $>8\%$	8.9%	—	8.5%	46.4%	35.6%
	Age ≤ 30	2.5%	—	0.9%	6.9%	4.5%
	Age 71+	43.8%	—	44.6%	29.9%	38.0%
	Race: African Am.	18.1%	—	22.0%	28.7%	19.0%
	Insulin: none	49.0%	—	42.3%	27.4%	40.9%

Table 3: Demographic composition of $p = 5\%$ victim sets versus training-set base rates. **Bold** indicates statistically significant enrichment or depletion at $p < 0.001$. TabDDPM excluded from Diabetes 130-US; see Section 4.2

composed of defaulters, but the specific individuals at risk are drawn from qualitatively different socioeconomic profiles.

Diabetes 130-US: HbA1c monitoring status is the dominant axis; age enrichment inverts by model class. HbA1c measurement status produces the strongest demographic signal on this dataset. TabSyn’s victim set closely mirrors the training distribution: 82.5% of its $p = 5\%$ victims have no HbA1c test recorded, nearly identical to the base rate of 81.7%. GReaT and Tabula invert this entirely: only 33.5% and 30.9% of their victims respectively are from the untested majority, despite that group comprising 81.7% of training records. For the HbA1c $> 8\%$ subgroup, GReaT’s victim rate reaches 46.4% against a base rate of 8.9% (+423%, chi-square $p < 10^{-300}$), and Tabula’s reaches 35.6% (+302%). Neither effect appears in TabSyn’s victim set. A hospital auditing this dataset using TabSyn would find no HbA1c signal and conclude that actively monitored patients face no elevated risk, a conclusion that is directly inverted under GReaT and Tabula.

Age enrichment shows a class-specific inversion. GReaT over-exposes patients aged ≤ 30 by +173% at $p = 5\%$ (6.9% vs. 2.5% base, $p = 1.3 \times 10^{-37}$) while under-exposing patients aged 71+ (−32%). Tabula shows the same directional pattern at lower magnitude. TabSyn inverts this: the ≤ 30 group is under-represented (−66%) and the 71+ group approximates the base rate. Young diabetic patients, 2.5% of training data but nearly 7% of GReaT’s $p = 5\%$ victims, are likely Type 1 patients with distinctive, frequently monitored clinical profiles that LLM-based models fit more tightly. GReaT additionally shows significant African American over-representation (28.7% vs. 18.1% base, +59%, $p = 6.4 \times 10^{-20}$).

As on UCI, the exclusive victim sets make the divergence concrete. LLM-exclusive victims show extreme HbA1c $> 8\%$ enrichment combined with strong ≤ 30 age over-representation. Diffusion-exclusive victims show moderate HbA1c enrichment and a broader age distribution. LLM-based models concentrate memorization risk on the most

metabolically active, frequently monitored patients; diffusion models distribute it more broadly. In both domains, the individuals most at risk differ substantially depending on which model is deployed.

5 Discussion

Our results establish that memorization risk in tabular synthetic data is not a scalar property of a model or dataset. It is a distribution over individuals, and that distribution shifts depending on which generative model is deployed. The generators we evaluate expose substantially different sets of training records, in some cases nearly disjoint, and the demographic composition of those sets differs in ways that are clinically and socioeconomically meaningful. On UCI, diffusion models concentrate risk on lower-income, lower-education defaulters while GReaT concentrates it on higher-income, higher-education defaulters. Both patterns carry harm, through different mechanisms and on different populations. We make no claim that one distribution of exposure is preferable to the other. Our contribution is narrower: this choice is currently being made without awareness that a choice is being made at all. On Diabetes 130-US, LLM-based models predominantly expose patients who have had HbA1c measured, a clinically engaged subpopulation that diffusion models largely do not touch. HbA1c is a marker of long-term blood glucose control, and patients tested for it tend to have more complete and more sensitive clinical records, meaning more information is at stake when their exposure occurs. In both cases, the individuals most at risk differ substantially depending on which model a practitioner selects.

This has a direct implication for how model selection should be understood. Choosing between TabSyn and GReaT to publish synthetic data is not a neutral technical decision about output quality or training efficiency. It is implicitly a decision about which subpopulation bears the privacy burden of that generation. This framing holds even when aggregate metrics suggest one model is safer than another: on UCI, GReaT’s lower DCR and exposure rate could lead an

auditor to conclude it poses less privacy risk. The per-sample analysis complicates this picture. GReaT does not uniformly reduce risk; it shifts the composition of who is exposed. A practitioner who selects GReaT on the basis of aggregate metrics has made a distributional choice without knowing it.

The Tabula findings directly illustrate this point. On UCI, Tabula’s victim overlap with diffusion models is nearly as high as diffusion-diffusion overlap, placing it closer to the diffusion class than to GReaT. On Diabetes 130-US, the pattern reverses: Tabula clusters tightly with GReaT and the cross-class gap is an order of magnitude larger than on UCI. This dataset-dependent behavioral shift means that architectural class alone is an unreliable guide to which individuals a model will expose. Two models from the same class can expose different people; two models from different classes can expose similar people depending on the dataset. The implication is that the audit problem is harder than knowing which architecture is deployed. It requires per-model, per-dataset evaluation of victim set composition, not class-level generalizations. A practitioner who selects a model based on architectural class alone may expose a substantially different subpopulation than a model from the same class would on a different dataset.

The demographic findings also surface a structural issue with how privacy audits are currently designed. Standard aggregate metrics are computed at the output level: they ask how close synthetic records are to training records on average. But the question of who bears risk is an input-level question about which training records are most exposed. These are not the same question, and optimizing for one does not guarantee anything about the other. Any model useful enough will expose some tail of individuals; the composition of that tail is a distributional choice that aggregate metrics do not surface. A model can have favorable aggregate metrics while concentrating exposure on a specific demographic subgroup, as GReaT demonstrates on UCI. An auditor who signs off on GReaT based on its exposure rate has cleared the model that produces the strongest demographic skew in our analysis. Conversely, a model with worse aggregate metrics may distribute exposure more broadly rather than concentrating it. Neither aggregate-level safety nor aggregate-level risk tells a practitioner what they need to know about distributional fairness.

Taken together, these findings suggest that privacy evaluation for synthetic data generation should be understood as a fairness problem, not just a security problem. The question is not only how much leakage exists but who bears it, and whether that burden falls equitably across the population or concentrates on individuals who are already vulnerable by virtue of their clinical profiles, financial circumstances, or demographic characteristics. Institutions generating synthetic data from sensitive domains should treat model selection as a decision with distributional consequences, and should evaluate candidate models not only on aggregate privacy metrics but on the demographic composition of their victim sets. An institution that selects a model on utility or aggregate privacy metrics alone has made that distributional decision without knowing it was being made. Aggregate metrics are a necessary but insufficient basis for claim-

ing that a synthetic dataset is safe to publish.

6 Limitations

This analysis uses Gower distance with equal feature weighting, treating all columns as equally relevant to memorization risk. Features that are more privacy-sensitive, such as diagnosis codes or payment history, may warrant higher weight, and a distance metric that reflects differential feature sensitivity could shift victim set composition. Developing privacy-aware distance metrics for mixed-type tabular data is a direction for future work.

The findings are established on two datasets from financial and clinical domains. While the replication across domains with different feature structures and demographic compositions is encouraging, two datasets cannot establish universality. Whether the architecture-dependent memorization patterns observed here generalize to other domains, dataset sizes, or feature types remains an open question.

On Diabetes 130-US, the diffusion class is represented by a single model (TabSyn), as TabDDPM fails architecturally on high-cardinality ICD-9 features. If TabSyn is atypical among diffusion models, some D130 findings may not generalize to the diffusion class as a whole. A broader evaluation including additional diffusion architectures would strengthen the cross-class claims on this dataset.

Memorization scores are proximity-based signals, not confirmed privacy breaches. A training record with high `mem_score` has synthetic neighbors close to it in Gower distance, indicating that the generative model’s distribution is concentrated near that record. This does not establish that the record can be extracted by an adversary or that a specific attack would succeed. The scores should be interpreted as privacy risk indicators rather than confirmed leakage.

The demographic categories used in this analysis are coarse: four age brackets, four credit limit bins, broad racial and education groupings. Finer-grained demographic analysis might reveal more localized exposure patterns or resolve cases where group-level enrichment is diluted by within-group heterogeneity. Small demographic subgroups in particular may face concentrated exposure that aggregate enrichment ratios do not fully capture.

Finally, a mechanistic investigation into why specific demographic subgroups are disproportionately exposed under each architecture is out of scope for this paper, which is primarily empirical; we treat it as a direction for future work.

7 Conclusion

Across two datasets from sensitive domains, we find that the choice of generative model for synthetic data generation is not privacy-neutral. Different tabular generators expose substantially different sets of training records, and the demographic subgroups most at risk shift substantially depending on which model is deployed. The pattern of overlap is not reliably predicted by architectural class alone: some within-class pairs expose largely the same individuals while others diverge sharply, depending on the dataset. Standard aggregate privacy metrics fail to surface this divergence, and in some cases can produce orderings that conflict

with subgroup-level findings: a model that performs better on aggregate measures may shift the composition of who is exposed rather than uniformly reducing risk.

The central implication is that model selection for synthetic data generation should be treated as a fairness decision, not only a technical one. When institutions choose a generative model without evaluating the demographic composition of its victim set, they are implicitly deciding who bears the privacy burden of that generation, without awareness that the decision is being made at all. Per-sample, per-model evaluation on the specific dataset being generated is necessary to make that decision.

Future work should extend this analysis to additional model architectures, domains, and dataset sizes, and examine whether differential privacy mechanisms reduce demographic skew in victim sets or merely shift it. A deeper mechanistic investigation into why specific demographic subgroups are disproportionately exposed under each model would complement the empirical patterns documented here: the generation biases and distributional failure modes that produce demographic skew are not yet fully understood, and understanding them is necessary for designing generators that distribute privacy risk more equitably. More broadly, the field needs evaluation frameworks that treat privacy protection as a distributional property, asking not only how much leakage exists but whether its burden is equitably distributed across the populations that sensitive data is meant to serve.

Ethical Statement

This work relies exclusively on two publicly available, de-identified benchmark datasets (UCI Default of Credit Card Clients and Diabetes 130-US) that are already stripped of direct identifiers; no new individual-level privacy exposure is created by our experiments, and no real individuals are re-identified or contacted as part of this research.

Our central finding, that memorization risk concentrates on identifiable demographic subgroups and that this concentration shifts depending on which generative model is deployed, is intended to inform more equitable practice, not to enable attacks. Knowing that a model class over-exposes a particular subpopulation is a prerequisite for institutions to correct for it, whether through model selection, differential privacy, or targeted mitigation. We recognize a dual-use consideration: the same per-sample memorization signal that lets a practitioner audit a model for fairness could, in principle, help an adversary identify likely victims under a known deployed model. We judge the balance to favor disclosure, since the aggregate metrics currently used in practice already fail to protect these subpopulations and leave them invisible to defenders while remaining exploitable by anyone with dataset and model access; withholding the analysis would not remove the underlying risk, only the ability to detect and mitigate it.

We also note that our demographic categories (e.g., income bracket, education level, race, age group) are coarse groupings drawn from the datasets' existing fields, used here to characterize aggregate exposure patterns rather than to make claims about any individual. Readers should not treat enrichment ratios for a demographic category as a statement

about any single person's risk. Finally, we intend our broader argument, that model selection for synthetic data is implicitly a fairness decision, to encourage institutions publishing synthetic data in sensitive domains to adopt subgroup-aware privacy evaluation rather than relying on aggregate metrics alone.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant 402042.

References

- Annamalai, M. S. M. S.; Gadotti, A.; and Rocher, L. 2024. A Linear Reconstruction Approach for Attribute Inference Attacks against Synthetic Data. In *Proceedings of the USENIX Security Symposium*.
- Borisov, V.; Seßler, K.; Leemann, T.; Pawelczyk, M.; and Kasneci, G. 2023. Language Models are Realistic Tabular Data Generators. In *Proceedings of the International Conference on Learning Representations*.
- Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*.
- Carlini, N.; Hayes, J.; Nasr, M.; Jagielski, M.; Sehwag, V.; Tramer, F.; Balle, B.; Ippolito, D.; and Wallace, E. 2023. Extracting training data from diffusion models. In *Proceedings of the USENIX Security Symposium*.
- Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; and Erlingsson, U. 2021. Extracting training data from large language models. In *Proceedings of the USENIX Security Symposium*.
- Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. *Advances in Neural Information Processing Systems*.
- Dwork, C. 2008. Differential Privacy: A Survey of Results. In Agrawal, M.; Du, D.; Duan, Z.; and Li, A., eds., *Theory and Applications of Models of Computation*, volume 4978, 1–19. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-79227-7 978-3-540-79228-4. Series Title: Lecture Notes in Computer Science.
- Dwork, C.; and Roth, A. 2014. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3-4): 211–487.
- Fang, Z.; Jiang, Z.; Chen, H.; Li, X.; and Li, J. 2025a. Understanding and Mitigating Memorization in Diffusion Models for Tabular Data. In *Proceedings of the International Conference on Machine Learning*.
- Fang, Z.; Jiang, Z.; Chen, H.; Zhang, X.; Tang, K.; Li, X.; and Li, J. 2025b. A Closer Look on Memorization in Tabular Diffusion Model: A Data-Centric Perspective. ArXiv:2505.22322 [cs].
- Giuffrè, M.; and Shung, D. L. 2023. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digital Medicine*, 6(1): 186.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*.

Hu, H.; Salicic, Z.; Sun, L.; Dobbie, G.; Yu, P. S.; and Zhang, X. 2022. Membership Inference Attacks on Machine Learning: A Survey. *ACM Computing Surveys*, 54(11s): 1–37.

Jordon, J.; Szpruch, L.; Houssiau, F.; Bottarelli, M.; Cherubin, G.; Maple, C.; Cohen, S. N.; and Weller, A. 2022. Synthetic Data – what, why and how? *ArXiv:2205.03257 [cs]*.

Kotelnikov, A.; Baranchuk, D.; Rubachev, I.; and Babenko, A. 2023. TabDDPM: Modelling Tabular Data with Diffusion Models. In *Proceedings of the International Conference on Machine Learning*.

Meeus, M.; Guepin, F.; Crețu, A.-M.; and De Montjoye, Y.-A. 2024. Achilles’ Heels: Vulnerable Record Identification in Synthetic Data Publishing. In Tsudik, G.; Conti, M.; Liang, K.; and Smaragdakis, G., eds., *Computer Security – ESORICS 2023*, volume 14345, 380–399. Cham: Springer Nature Switzerland. ISBN 978-3-031-51475-3 978-3-031-51476-0.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *Proceedings of the IEEE Symposium on Security and Privacy*. IEEE.

Somepalli, G.; Singla, V.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2023. Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Stadler, T.; Oprisanu, B.; and Troncoso, C. 2022. Synthetic Data - Anonymisation Groundhog Day. In *Proceedings of the USENIX Security Symposium*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems*.

Ward, J.; Gu, B.; Wang, C.-H.; and Cheng, G. 2025. When Tables Leak: Attacking String Memorization in LLM-Based Tabular Data Generation. *arXiv preprint arXiv:2512.08875*.

Ward, J.; Wang, C.-H.; and Cheng, G. 2024. Data Plagiarism Index: Characterizing the Privacy Risk of Data-Copying in Tabular Generative Models. *ArXiv:2406.13012 [cs.LG]*.

Wiatrak, M.; Albrecht, S. V.; and Nystrom, A. 2020. Stabilizing Generative Adversarial Networks: A Survey. *ArXiv:1910.00927 [cs]*.

Zhang, H.; Zhang, J.; Shen, Z.; Srinivasan, B.; Qin, X.; Faloutsos, C.; Rangwala, H.; and Karypis, G. 2024. Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space. In *Proceedings of the International Conference on Learning Representations*.

Zhao, Z.; Birke, R.; and Chen, L. Y. 2025. TabuLa: Harnessing Language Models for Tabular Data Synthesis. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*.