SYNTHETIC DATA SHIELD: OBFUSCATING CONTEXTS FOR IN-CONTEXT TABULAR PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Tabular foundation models are becoming a popular approach to tabular classification based on in-context prediction. To classify a new entry, the model receives the labeled training set and the new row, then predicts its label in a single inference pass without task-specific training on the target dataset. A major privacy concern of this paradigm is that when a third party performs classification, the training data used as context must also be shared with that party. We propose **Synthetic Data Shield** (SDS), a method that does not require sending real training data to a third party. SDS creates a local synthetic training context and sends only this synthetic data, keeping the original training records hidden. The foundation model remains unchanged, while accuracy stays close to that of the original data. We evaluate our method through three privacy axes: membership inference, near-record exposure, and attribute inference. Our analysis shows that prediction utility, membership leakage, near-record exposure, and attribute inference relate to different statistical properties of the synthetic data generator, allowing us to control membership and near-record exposure separately while attribute inference tracks model accuracy in our measurements. Membership attacks stay close to chance, and synthetic records sit nearly as far from the training rows as from held-out rows of the same population. Attribute inference differs because statistical relations useful for prediction can also help predict a hidden attribute. We further show that adding noise can reduce this leakage but also causes a small loss in model accuracy. We test SDS on 71 tabular datasets and across tabular foundation models such as TabPFN-2.5, TabICL, LimiX, TabFM, and TabuLa-8B. We use the same synthetic contexts across these models (TabuLa-8B reads a 32-shot subset of each), with the mean accuracy loss staying within 2.5 percentage points for the tested transfer models.

1 INTRODUCTION

Tabular foundation models (TFMs) such as TabPFN (Hollmann et al., 2023; 2025; Grinsztajn et al., 2025), TabICL (Qu et al., 2025; 2026), LimiX (LimiX Team, 2025), Google's TabFM (Google Research, 2026), and the LLM-based TabuLa-8B (Gardner et al., 2024) have emerged over the last few years as dominant classifiers for tabular data, changing how we work with it. They are in-context predictors that read the labeled training set as context and predict new examples in a single forward pass. They achieve state-of-the-art accuracy on datasets with up to several hundred thousand rows and a moderate number of classes. This convenience carries a cost. When we delegate the classification of new rows to a third party or hosted service, we must share the training data for context, creating a privacy bottleneck.

This exposure is not rare. Consider a hospital that uses an externally hosted TFM service for clinical-risk prediction or a bank that deploys a centrally hosted in-context model for credit scoring. In both cases, they must reveal the training rows that serve as context. Such a third party may be honest, in the sense that it runs the requested inference exactly, but it also observes the full context and may store or analyze it. We therefore focus on an honest-but-curious prediction service and ask whether the data owner can replace the real context with a synthetic one while preserving the TFM's accuracy.

Differential privacy (DP) (Dwork et al., 2006; Dwork & Roth, 2014) is the natural first candidate, a framework that protects individual records by adding carefully calibrated random noise to the training data and its derivatives. In our case, injected noise erases the signal TFMs rely on, and our

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

experiments show accuracy losses above 40 points for a practical DP mechanism at the budgets that give meaningful privacy (App. L.14). Ordinary tabular synthesizers are models fitted to a table to generate new artificial rows with similar statistics, and most lose accuracy in the small-context regime, while the strongest of them, ForestDiffusion, keeps accuracy but leaks membership. Deep generators such as CTGAN/TVAE, TabDDPM, GOGGLE, and ForestDiffusion (Xu et al., 2019; Kotelnikov et al., 2023; Liu et al., 2023; Jolicoeur-Martineau et al., 2024) are powerful but require many rows per class and can memorize the training data. DP synthesizers such as PrivBayes and MST (Zhang et al., 2017; McKenna et al., 2021) discretize continuous features by converting continuous values into coarse bins, losing useful signal. Lastly, geometry-preserving encodings such as keyed random Fourier features (K-RFF) (Rahimi & Recht, 2007) scramble each record using a secret key and can preserve accuracy while maintaining the pairwise distances between records so the geometry of the original records remains recoverable.

Our approach rests on a single observation: prediction utility and three leakage modes each depend on a different generator property, so membership leakage and near-record exposure can be reduced with separate levers while attribute inference stays tied to utility. Building on this, we propose Synthetic Data Shield (SDS). In SDS, the data owner locally fits a small per-class generator, samples a synthetic labeled context, and sends only this synthetic context to the prediction service, which then runs the same pre-trained in-context model on it. No real training record leaves the data owner's machine. SDS is not a replacement for DP and does not provide the uniform per-record guarantee that DP offers. It is, rather, an obfuscation layer for the practical deployment setting in which the prediction service should not see the original records and in which the utility must remain close to the TFM's real-context utility. One kind of leakage remains as a trade-off. Attribute inference, in which an attacker predicts a hidden column of a real record, succeeds at $R^2 \approx 0.77$ on the three real-data core datasets, where R^2 is a prediction-quality score defined in §2. For deployments where this is unacceptable (e.g., regulated demographic columns), we suggest an optional hardening step (§6) that reduces attribute leakage to $R^2 \approx 0.52$ at a $\sim 1\%$ utility cost.

Our contributions are as follows:

- We formulate privacy-preserving delegated inference for in-context tabular prediction, where the sensitive object is the labeled context itself, and introduce SDS, which is a simple type-aware synthetic-context pipeline for numeric, mixed-type, and mostly-categorical tables.
- We analyze utility, membership leakage, attribute inference, and near-record exposure using four distributional quantities: Wasserstein distance, replace-one KL divergence, total variation, and generator geometry (§5).
- We evaluate SDS across 71 datasets in five suites (106 evaluations) plus multiple attacks. We show that accuracy remains close to real-context accuracy, membership remains near chance, and synthetic records sit nearly as far from training rows as from held-out rows. These results hold across leading current-generation foundation models (TabPFN-2.5, TabICLv2, LimiX, Google's TabFM) and an 8-billion-parameter LLM-based predictor (TabuLa-8B), all with mean gaps within 2.5 pp.

2 Problem Formulation and Threat Model

We aim to use externally hosted in-context tabular foundation models as classifiers without exposing the sensitive training records. TFMs, built on prior-data fitted networks (PFNs) (Müller et al., 2022), are pre-trained on many synthetic datasets so that, at test time, they predict directly from a new dataset used as context. A TFM $q_\theta(y_{\text{test}} \mid x_{\text{test}}, \mathcal{D}_{\text{train}})$ with parameters θ predicts the label distribution at a test point from a labeled training context $\mathcal{D}_{\text{train}}$ in a single forward pass. This in-context learning (ICL) needs no re-training. Given a prior p over datasets $\mathcal{D} = (\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}})$, TFMs are pre-trained to minimize

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathcal{D}\sim p(\cdot)}\big[-\log q_\theta(y_{\text{test}} \mid x_{\text{test}}, \mathcal{D}_{\text{train}})\big].$$

The data owner's local labeled training set $\{(x_i, y_i)\}_{i=1}^N$, with $x_i \in \mathbb{R}^d$ and $y_i \in [C] = \{1, \ldots, C\}$, plays the role of $\mathcal{D}_{\text{train}}$. As the test split is unused, we use $\mathcal{D}$ for this training set. We abbreviate $f_\theta(x^*; X, y) := q_\theta(\cdot \mid x^*, (X, y))$ to be the predicted class distribution at query x^* given context (X, y) where X stacks the x_i as rows and y represents the labels.

Transmitting $\mathcal{D}$ directly exposes every record. To resolve this, we propose a **release frame** in which the data owner locally fits a class-conditional generator p_G on $\mathcal{D}$ (so $p_G = p_{G,\mathcal{D}}$ depends on the fitting data, and we suppress the subscript), models each class's features separately, samples a synthetic context $(\widetilde{X}, \widetilde{y})$ of size $m = \alpha N$ (the oversampling factor α sets synthetic size relative to real and $\alpha = 1$ provides equal size), and transmits *only* this context. Here *release* is used in the data-privacy sense, meaning the data is sent to another party. We denote $Q_{\mathcal{D}}$ as the release law for the distribution of the released context. The external model then computes $f_\theta(x^*; \widetilde{X}, \widetilde{y})$.

We operate under the standard **honest-but-curious** threat model (Goldreich, 2004; Lindell & Pinkas, 2009). The external service executes inference faithfully but may store and analyze all transmitted synthetic data and intermediate state. SDS protects the historical training corpus. The query rows sent for scoring still reach the service in the clear, and are out of scope: a record submitted for prediction is one the data holder is actively acting upon, typically under a live consent relationship, whereas the training cohort is not. We exclude active adversaries that inject malicious queries, modify inference, or collude. An adversary probing the hosted model's context attacks the service's own interface. Since the honest-but-curious service already holds the full released context, such probing adds nothing beyond the query-adaptive attack we already audit (§6).

For evaluation, we use one utility metric, measured on a test set T, and three privacy attacks.

The utility gap Δ_{util} is the change in top-1 accuracy from running on synthetic context instead of a real one. Under class imbalance, we use class-balanced accuracy (per-class accuracy averaged so rare classes count equally) such that

$$\Delta_{\text{util}} = \text{Acc}_T(f_\theta; \widetilde{X}, \widetilde{y}) - \text{Acc}_T(f_\theta; X, y). \tag{1}$$

The membership-inference advantage Adv_{MIA} is an attacker's ability to distinguish training from holdout records by their distance to a synthetic context. Following (Stadler et al., 2022), we score each record x by $s(x) = -\min_j \|x - \widetilde{x}_j\|_2$ (negated distance to the nearest synthetic record) and take the area under the ROC curve (AUC) of s separating training from matched holdout records (same source, unused). We report the chance-centered advantage

$$\text{Adv}_{\text{MIA}} = \max(\text{AUC}_s, 1 - \text{AUC}_s) - \tfrac{1}{2}. \tag{2}$$

Geometrical reconstruction R^2_{geom} measures how well an attacker reconstructs the spatial layout of the original records from synthetic pairwise distances (R^2=1 perfect, R^2=0 none). We embed the synthetic pairwise-distance matrix by classical multidimensional scaling (MDS, Cox & Cox, 2000), then align the embedded points to X by Procrustes (Gower, 1975), and report

$$R^2_{\text{geom}} = R^2\big(\text{Procrustes}(\text{MDS}(D_{\widetilde{X}})) \to X\big), \quad (D_X)_{ij} = \|x_i - x_j\|_2. \tag{3}$$

Here $R^2(A \to B)$ is the coefficient of determination when A predicts B, and $D_X, D_{\widetilde{X}}$ are the pairwise-distance matrices of the real and synthetic records. Because this index-paired score does not separate a synthetic release from a row-permuted copy of the real records (App. L.10), we retain it only as a control and assess the axis by the near-record rate, the fraction of training records with a synthetic neighbour within 0.5 standardized units.

Attribute inference R^2_{attr} measures how well a model trained solely on synthetic data predicts a sensitive column of a real record from its remaining columns. For feature j, we train g_j on synthetic data and evaluate on real records,

$$R^2_{\text{attr}}(j) = R^2\big(g_j(X_{-j}) \to X_j\big), \qquad g_j \in \arg\min_{g \in \mathcal{H}} \mathbb{E}_{Q_{\mathcal{D}}}(g(\widetilde{X}_{-j}) - \widetilde{X}_j)^2. \tag{4}$$

Here $\mathcal{H}$ is the attacker's model class, with ridge regression, a linear predictor with a penalty on the size of its coefficients, for numeric attributes and a gradient-boosted tree classifier for categorical ones.

Our formal objective is to design a release law $Q_{\mathcal{D}}$ that minimizes combined adversarial advantage across membership and attribute attacks, subject to a tolerance τ on accuracy degradation:

$$\min_{Q_{\mathcal{D}}} \quad \lambda_1 \,\text{Adv}_{\text{MIA}} + \lambda_3 \, R^2_{\text{attr}} \quad \text{s.t.} \quad \Delta_{\text{util}} \geq -\tau. \tag{5}$$

The non-negative weights λ_1, λ_3 set the relative cost of the two leakage types that enter the objective (near-record exposure is audited directly, App. L.10, since its index-paired reconstruction score is

not a valid privacy measure) and depend on the application, so a deployment that must protect a regulated attribute raises λ_3. We set both to 1 since we target no particular application, and we report the three leakage axes separately so readers with different priorities can reweigh them. We choose the tolerance τ the same way, and rather than commit to one value, we report Δ_{util} itself. Eq. (5) therefore states what any release law must trade off, and it is not a program that SDS solves. SDS is the fixed pipeline of §4, and our experiments in §6 measure its three leakage axes and its accuracy cost.

3 Related Work

In-context tabular models predict from labeled examples in one forward pass without per-dataset training. Prior-data fitted networks train a transformer on synthetic tasks to approximate the Bayesian answer for a context (Müller et al., 2022), and TabPFN applies this to tabular classification (Hollmann et al., 2023; 2025). We test SDS on the current models with this design: TabPFN-2.5 (Grinsztajn et al., 2025), TabICL/v2 (Qu et al., 2025; 2026), LimiX (LimiX Team, 2025), Google's TabFM (Google Research, 2026), and the 8B LLM-based TabuLa-8B (Gardner et al., 2024) (App. A).

Synthetic tabular data shares statistical structure without releasing the real rows (Rubin, 1993; Jordon et al., 2022), and privacy audits test whether those rows leak who was in the data, hidden attributes, or record layout (Stadler et al., 2022; van Breugel et al., 2023). We reuse these audits for a different question: whether a synthetic context preserves a pre-trained model's predictions.

Competing protections each give something up. Differential-privacy methods add noise for the strongest guarantee but cut TabPFN accuracy sharply, since PrivBayes, MST (Zhang et al., 2017; McKenna et al., 2021) and AIM (McKenna et al., 2022) coarsen continuous values into bins. Deep generators (CTGAN/TVAE, TabDDPM, GOGGLE, ForestDiffusion) and the DP generator PATE-GAN (Xu et al., 2019; Kotelnikov et al., 2023; Liu et al., 2023; Jolicoeur-Martineau et al., 2024; Jordon et al., 2019) need large datasets and can copy real records verbatim (Somepalli et al., 2023). The closest option built to preserve accuracy, a keyed random-Fourier-feature encoding (K-RFF, App. F.4) (Rahimi & Recht, 2007), scrambles rows with a secret key but keeps enough distance between rows to recover their layout. SDS needs no key. Cryptographic tools (enclaves, multi-party computation, homomorphic encryption) are stronger but too slow at full TabPFN context sizes (Hao et al., 2022).

Privacy-preserving in-context learning mostly protects examples in a language-model prompt (Tang et al., 2024; Wu et al., 2024; Duan et al., 2023; Hong et al., 2024). Nearest to us, DP-TabICL (Carey et al., 2024) and TabPATE (Wahdany et al., 2026) add differential privacy to *tabular* in-context learning at a DP accuracy cost. SDS is complementary and uses no DP. It swaps the real context for a synthetic one, defending the distances, attributes, and row-level membership that structured rows expose (App. A).

4 Synthetic Data Shield

To address the privacy problem formulated in §2, we present the *Synthetic Data Shield* (SDS), a local pre-processing layer that replaces the training set with synthetic data for in-context prediction. In this framework, generation is class-conditional. A generator is fitted per label, and each synthetic row is emitted with that label so the context arrives already labeled. Context size is often only tens to hundreds of rows per class, so the generator stays small and stable. SDS is type-aware, meaning numeric, mixed, and mostly categorical tables need different modeling assumptions, as described below.

The pipeline and its data flow appear in Algorithm 1 and Figure 1. The SDS algorithm examines column types, routes the table to one of three generators, and then samples K candidate contexts of size m (K=1 by default). Optional selection keeps, among the K candidates, the one whose context best re-predicts the owner's real labels (*self-prediction*), which discards batches dominated by tail-sampled outliers. An optional distance-to-closest-record (DCR) filter is applied, and the resulting context is transmitted.

For numeric tables, deep generators overfit at small N_c (Table 9) and lose accuracy in the small-context regime (Table 15), so for each class c SDS fits a 50:50 ensemble of two density models, a

Algorithm 1 Synthetic Data Shield (SDS)

Require: Local dataset $\mathcal{D} = (X_{\text{real}}, y)$; oversampling factor α (set to 2 in all main experiments); number of candidates K (K=1 in all headline runs, which skip selection); optional DCR threshold τ_{DCR}.

Ensure: Synthetic context $(X_{\text{synth}}, y_{\text{synth}})$.

$\rho_{\text{cat}} \leftarrow$ fraction of categorical columns in X_{real}
if $\rho_{\text{cat}} > 0.8$ and at most two numeric columns **then**
    Fit a class-conditional Chow–Liu (Chow & Liu, 1968) generator with feature flips and label refinement. ▷ see "Mostly categorical tables"
else if $\rho_{\text{cat}} > 0$ **then**
    Fit the mixed-type copula generator. ▷ see "Mixed-type tables"
else
    Fit the class-conditional equal-weight mixture of a BIC-selected GMM and a Gaussian copula ($\frac{1}{2}$BIC-GMM + $\frac{1}{2}$Gaussian-copula). ▷ see "Numeric tables"
end if
for $k = 1, \ldots, K$ **do**
    Sample $(X^{(k)}_{\text{synth}}, y^{(k)}_{\text{synth}})$ with $\lfloor \alpha N \rfloor$ rows, allocated across classes in proportion to the real class counts (a quota of αN_c rows for class c, where N_c is the number of real rows with label c).
end for
Select $k^* = \arg\max_k \text{acc}\big(f_\theta(X_{\text{real}}; X^{(k)}_{\text{synth}}, y^{(k)}_{\text{synth}}), y\big)$.
if τ_{DCR} is provided **then**
    For each row x of the selected context, let $\delta(x) = \min_i \|x - X_{\text{real},i}\|_2$; accept x with prob. $\min(1, (\delta(x)/\tau_{\text{DCR}})^2)$ ▷ smooth DCR rule; rejected rows are resampled until the per-class quota is filled, App. F.25
end if
return $(X^{(k^*)}_{\text{synth}}, y^{(k^*)}_{\text{synth}})$ to the prediction service.

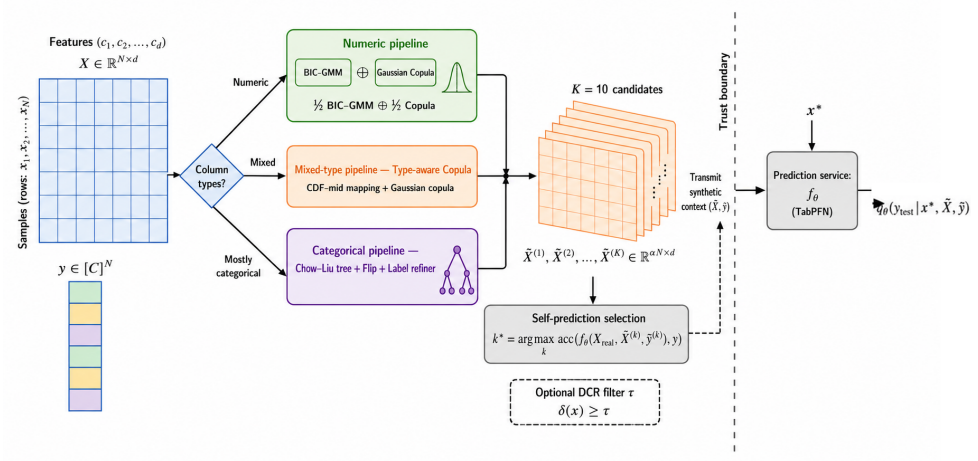

Figure 1: SDS data flow. A type-detection step routes the local context to one of three class-conditional synthesizers, then K candidates of shape $\alpha N \times d$ are sampled (K=1 by default), the best self-predicting one is selected when $K > 1$, an optional DCR filter rejects near-record points, and only the selected context crosses the trust boundary to reach the prediction service f_θ.

Gaussian mixture model (Bishop, 2006) and a Gaussian copula (Sklar, 1959) with class-conditional densities p_{GMM} and p_{Cop}, averaged into the ensemble density p_{ens} such that

$$p_{\text{ens}}(x \mid c) = \tfrac{1}{2} p_{\text{GMM}}(x \mid c) + \tfrac{1}{2} p_{\text{Cop}}(x \mid c).$$

The GMM component count is chosen by the Bayesian Information Criterion (BIC, Schwarz, 1978) with a small-data cap $k_{\max,c} = \max(1, \min(8, \lfloor N_c/(d+2) \rfloor))$. The copula estimates inter-column dependency with a Ledoit–Wolf regularized covariance that stays stable when $N_c < d$ (Ledoit & Wolf, 2004). Table 8 confirms the ensemble beats either component alone.

For mixed-type tables, one-hot encoding a categorical column adds a separate column per category, inflating the feature count d. SDS avoids this by handling numeric and categorical columns differently and modeling them jointly with a copula by keeping d small. At low categorical fraction, categorical columns are sampled from smoothed empirical class-conditional distributions and numeric columns within the corresponding subgroup. At higher fractions, each categorical value is instead placed at the midpoint of its (smoothed) slice of the column's empirical cumulative distribution function (CDF) such that a Gaussian copula can model it jointly with the numeric columns. For small datasets ($N < 400$), SDS falls back to ordinal encoding. This preserves both types' marginals without dimensional blow-up.

For mostly categorical tables, SDS fits a Chow–Liu tree per class (Chow & Liu, 1968) where each column links to one parent via a Laplace-smoothed conditional probability table (pseudo-counts keep unseen combinations nonzero). To reduce row copying, each generated value is replaced with probability $p_{\text{flip}} = 0.15$ by a draw from the class-conditional marginal. A locally trained gradient-boosted label refiner then predicts labels for the generated rows while correcting higher-order rules a pairwise tree misses. This keeps the nearest-neighbor membership attack of Eq. (2) (NN-MIA) at chance (Table 18) while preserving the joint structure that drives utility.

Selection is *optional* and off in all headline runs, which sample a single context ($K{=}1$). Enabling it ($K{=}10$) matches $K{=}1$ on utility and on every leakage axis within per-fold noise on our 14-dataset headline benchmark (Table 40, App. J.4). Because selection is data-dependent, we audit the selected context as well. It does not need the deployed model's weights either. Ranking the same 10 candidates with a local 100-tree Random Forest instead of the production TFM leaves the TFM's downstream accuracy unchanged to within 0.001, so we recommend the local proxy as the scorer. An optional DCR filter (App. F.25) probabilistically rejects synthetic points too close to real records, trading utility for tighter near-record control.

5 Analysis

Each of the four axes of §2 is governed by a distinct distributional quantity relating the generator, or its release, to real data, not by a single measure of how far p_G is from $p^\star$. Two assumptions recur below: **context-Lipschitz** of f_θ with constant L_f (Assumption 1, measured in App. E) and **GMM regularity** (BIC component counts stable on $\geq 77\%$ of replace-one perturbations, mean 95%, across the 14-dataset numeric tier, App. C.15). Theorems 1, 2 and 3 are sufficient distributional conditions, not tight numerical bounds. Proofs are in App. C, and App. C.1 details the three levels of support (theoretical, conditional, and empirical) for SDS's components.

5.1 Utility that is controlled by distributional fit

Theorem 1 links the prediction shift driving Δ_{util} in Eq. (1) to the Wasserstein-1 distance (Villani, 2008).

Theorem 1 (Utility stability via transport distance)**.** *Assume f_θ is context-Lipschitz (its prediction shifts by at most L_f times the context change). Let $\hat{\mu}_n$ be the empirical measure of n labeled synthetic samples (feature vectors with their labels) drawn from p_G, and let $\hat{\mu}_N$ be the empirical measure of the N real labeled records from $p^\star$. Then, with the expectation below taken over the random draws of both samples,*

$$\mathbb{E}\big[\|f_\theta(x_{\text{test}}; \hat{\mu}_n) - f_\theta(x_{\text{test}}; \hat{\mu}_N)\|_\infty\big] \leq L_f \left[\mathcal{W}_1(p_G, p^\star) + R_n + R_N\right],$$

where R_n, R_N are finite-sample terms that vanish as n and N grow (of order $n^{-1/d}$ and $N^{-1/d}$, explicit form in App. C.3).

This motivates the SDS generator, which fits each class's distribution accurately enough to preserve predictions but simply enough to avoid memorizing rows. The GMM/copula ensemble strikes this balance with a KL divergence no worse than the average of its two components (App. C.10). The Lipschitz assumption is plausible because f_θ depends on its context only through attention, a bounded aggregation over the context rows, which limits but does not by itself dilute any single row's influence (attention can concentrate on one row), and the measured sensitivity is moderate (L_f median 20.3, 95th percentile 112.8, App. E). We use Theorem 1 structurally: it identifies $W_1(p_G, p^\star)$ as the

operative quantity for utility but is one-sided and does not predict loss size, consistent with the weak cross-dataset W_1–Δ_{util} correlation (Pearson r=0.07, p=0.80, App. F.21).

5.2 Membership is controlled by replace-one stability

Membership attacks become possible when replacing one record noticeably changes the release distribution. SDS does not satisfy DP, but the same replace-one view identifies the quantity to track.

Theorem 2 (Replace-one divergence controls membership advantage)**.** *Let $Q_{\mathcal{D}}$ be the release law of SDS, write $\mathcal{D} \simeq \mathcal{D}'$ for two datasets that differ in one feature vector and not in labels, and let*

$$\Gamma_\infty = \sup_{\mathcal{D}\simeq\mathcal{D}'} \max\{\mathrm{KL}(Q_{\mathcal{D}}\|Q_{\mathcal{D}'}), \mathrm{KL}(Q_{\mathcal{D}'}\|Q_{\mathcal{D}})\}.$$

Then every balanced replace-one membership adversary A (an attacker that observes the release and guesses, with equal prior, which of the two datasets in a pair $\mathcal{D} \simeq \mathcal{D}'$ produced it, and whose advantage $\mathrm{Adv}(A)$ is its accuracy above the chance level $\frac{1}{2}$) satisfies

$$\mathrm{Adv}(A) \le \sqrt{\Gamma_\infty/8}.$$

The theorem bounds an accuracy advantage, while Eq. (2) reports an AUC advantage, which is never smaller, so the theorem guides rather than bounds the audited number. Under GMM regularity (App. C.15), the theorem's $O(\sqrt{\alpha/N})$ Pinsker rate requires the per-record KL to shrink fast enough to offset the $m = \alpha N$ released records. Across the 14-dataset numeric tier, the per-record KL runs from 0.0008 to 5.22 nats, the unit of KL divergence under the natural logarithm, and it stays at or below 0.05 on eight of the fourteen datasets, with a median of 0.03 (Table 5). The audited advantage stays at 0.01–0.10 (Table 9). The bound is numerically vacuous under the oversampled regime of §6, but the corresponding diagnostic becomes numerically non-vacuous under the capped-release regime SDS uses for large datasets (Waveform $1.02 \to 0.145$ against an audited $\approx$0.05, App. L.8).

5.3 Reconstruction procedures converge to generator geometry

Proposition 1 decouples the reconstructible geometry from the empirical record geometry driving R^2_{geom} in Eq. (3).

Proposition 1 (Generator-geometry convergence)**.** *Let $x_i^s \overset{\text{iid}}{\sim} p_G^X$, the feature marginal of the generator (the distribution of a synthetic feature vector with its label ignored), which has finite second moment. Continuous geometrical-reconstruction procedures such as MDS or principal component analysis (PCA) applied to $\{x_i^s\}$ converge almost surely to the population geometry of p_G^X (its distribution of pairwise distances), which generically (for all but degenerate configurations) differs from the empirical geometry of the training set when p_G^X is a k-component mixture with $k < N$. This concerns the limiting distance distribution only: it does not imply that individual training records cannot be recovered, since a generator can retain record-level information even when its population geometry differs from the empirical one.*

5.4 Attribute inference is a structural residual risk

Preserving the joint structure of the real distribution also preserves predictable relationships among attributes, which helps prediction but becomes a source of attribute-inference risk. Theorem 3 bounds the *excess* risk of the synthetic-trained attribute predictor that underlies R^2_{attr} in Eq. (4), its suboptimality relative to the best real-data predictor, by $2\,\mathrm{TV}(p^\star, p_G)$.

Theorem 3 (Attribute-inference transfer)**.** *Let $\ell \in [0,1]$ be a loss function and let R_P, R_G denote the corresponding risks (expected losses) under the real and generator distributions. For $h_G \in \arg\min_{h\in\mathcal{H}} R_G(h)$, where $\mathcal{H}$ is the class of attribute predictors from Eq.* (4)*,*

$$R_P(h_G) \le \inf_{h\in\mathcal{H}} R_P(h) + 2\mathrm{TV}(p^\star, p_G).$$

Thus, a high-utility synthesizer can support attribute prediction not via memorization but by learning true population correlations. Empirically, SDS and ForestDiffusion reach similar attribute-inference performance despite SDS's much lower membership leakage (§6).

Theorem 3 assumes a bounded loss, the standard condition for a total-variation transfer bound, because such a loss's risk changes by at most its range times the TV distance. Our reported Ridge R^2 uses unbounded squared error and is therefore evaluated empirically, with a truncated-loss analogue in Remark 1.

6 Experiments

6.1 Setup

Experiments run on a Slurm cluster with $8\times$ NVIDIA B200 GPUs. Per-dataset results appear in App. F. We evaluate SDS on six in-context readers: TabPFN (our re-implementation of the TabPFN-2.5 weights (Grinsztajn et al., 2025)), the official TabPFN-2.5 package, TabICLv2 (Qu et al., 2026), LimiX (LimiX Team, 2025), TabFM (Google Research, 2026), and the 8B LLM TabuLa-8B (Gardner et al., 2024). All six run on the 14-tier under identical SDS releases (5-fold $\times$ 3 seeds), with TabPFN in Table 11 and the other five in Table 3. The full five-suite evaluation uses TabPFN and spans the 106 dataset–protocol pairs of Table 1, with per-dataset tables for the breadth and TabZilla-hard tiers in Tables 35 and 36. The tiers overlap, so these 106 pairs come from 71 *distinct* datasets (per-dataset provenance in App. A). Headline protocols sample a single context ($K{=}1$) at $\alpha{=}2$, and both choices are robust (App. J.2, J.4). Baselines include CTGAN/TVAE (Xu et al., 2019), ForestDiffusion (Jolicoeur-Martineau et al., 2024), PrivBayes (Zhang et al., 2017), K-RFF (App. F.4), and a Random Forest (Breiman, 2001).

SDS stays within a few points of real-context TabPFN accuracy on all five suites (Table 1), and on Wine its decision regions match the real context's (Fig. 2). On seven imbalanced binary datasets balanced accuracy moves by only -0.7 points, while minority precision and PR-AUC fall by 10.3 and 6.1, a limitation the headline metric hides (App. K). Confidence intervals and non-inferiority tests are in App. F.

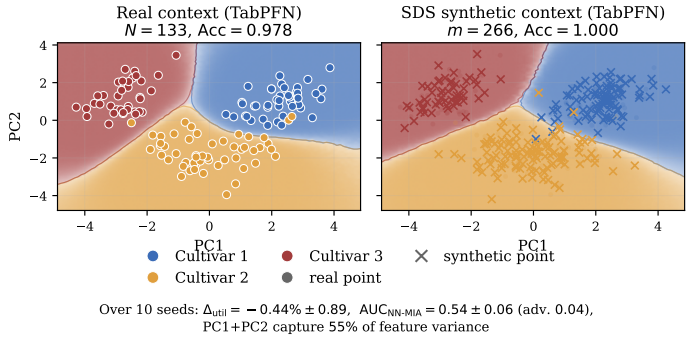

Figure 2: TabPFN's class-probability landscape on Wine ($N{=}178$, $d{=}13$, $C{=}3$) under one fold's 133-row real context (left) vs. its 266-row SDS synthetic context ($\alpha{=}2$, right). Decision regions are nearly identical while synthetic points cover the same class-conditional support without reproducing any real record.

Table 1: Utility across the five suites: mean accuracy gap Δ_{util} (SDS minus real) and datasets within $|\Delta| \leq 5\%$.

Suite	**#**	**Mean Δ_{util}**	**Within 5%**
Core six	6	-0.2%	6/6
14-tier (headline)	14	-1.4%	13/14
Mixed-type	15	-2.6%	13/15
Breadth	40	-2.7%	33/40
TabZilla hard	31	-3.9%	24/31

Table 2: Privacy leakage on the 14-tier (lower is better), baseline ForestDiffusion, with the attribute row on the real-data core. Geometry is omitted: the index-paired metric scores a row-permuted copy of the real records as it scores SDS (App. L.10).

Axis (metric)	**SDS**	**Base**
Utility (Δacc)	-1.4%	–
Membership (NN-MIA)	**0.54**	0.73
Attribute, real-data core	0.77	0.76

Table 3: Model transfer on the 14-tier: real vs. SDS mean accuracy, Δ = Acc SDS − Acc real (unrounded). †TabuLa-8B is an 8B LLM at 32-shot (App. L.7), where only the gap is of interest.

Model	Acc (real)	Acc (SDS)	Δ	Within 5%	Worst dataset
TabPFN-2.5	0.945	0.931	-1.3%	13/14	Vehicle, -6.7%
TabICLv2	0.944	0.930	-1.3%	13/14	Vehicle, -7.3%
LimiX-16M	0.940	0.929	-1.1%	13/14	Vehicle, -5.4%
TabFM	0.948	0.929	-1.9%	12/14	Vehicle, -10.6%
TabuLa-8B†	0.625	0.600	-2.5%	11/14	Pendigits, -11.3%

We compare SDS against existing synthesizers before turning to the three privacy axes. Among prior generators, ForestDiffusion comes closest on accuracy, but it leaks far more membership information (NN-MIA AUC 0.73 vs. 0.54). The remaining synthesizers lose substantial accuracy in the small-context regime (14–61% mean gaps, Tables 15 and 24), and AIM (McKenna et al., 2022) falls far below real-context accuracy at $\varepsilon \in \{1, 10\}$ (Table 25), a cost that, for the class-conditional Gaussian mechanism we control, comes from the privacy noise rather than the implementation, since that mechanism loses only 2.6 points at an effectively unbounded budget (App. L.13). Sweeping its budget shows it never dominates SDS on both axes at once (App. L.14). SDS with TabPFN still matches or exceeds a strong non-foundation baseline, a Random Forest, on $10/14$ datasets ($+1.0\%$ mean, Table 27).

SDS holds membership attacks near chance under every attack we run on the single-release headline audit (Table 2), and under the nearest-neighbor MIA (NN-MIA) of Stadler et al. (2022), its advantage is lower than ForestDiffusion's on all 14 datasets of the 14-tier (Table 9). Four stronger attacks stay near chance as well: the query-adaptive QAPD-MIA, a shadow-model attack (Shokri et al., 2017), the density-ratio DOMIAS (van Breugel et al., 2023), and the no-box local-likelihood-ratio Gen-LRA (Ward et al., 2025), which also separates SDS from a noise-perturbed release of the real rows at 0.55 against 0.71 (App. L.16). Averaging each dataset's strongest no-box attack gives AUC 0.56, whereas the same attacks reach 0.73 against a deliberately memorizing generator, confirming they are live rather than weak (App. F.19). Pooling out to 200 releases raises the attack by at most 0.075 AUC on twelve of fourteen but carries Iris to 0.75 and Seeds to 0.76 (App. L.15).

Geometry is the second axis. Reconstruction falls to near zero against K-RFF (Table 10), but that metric pairs rows by index and scores a row-permuted copy of the real records the same, so we rest the axis on near-record control instead (App. L.10). The optional DCR filter further tightens near-record control on low-d data, and distributional discriminators (logistic/MLP) also stay near chance on all but Segment (App. F).

Attribute inference is the one axis where the leakage is structural rather than near chance. A ridge attacker trained on SDS synthetic data and evaluated on real records reaches mean $R^2_{\text{attr}} = 0.59$ on the 14-tier (App. F.13) and ties ForestDiffusion on the real-data core (Table 2). The near tie is consistent with Theorem 3 under closeness in total variation. The operative quantity is the excess over an attacker who never sees the release. That excess is small and shrinks as the attacker strengthens, from 0.087 with Ridge to 0.021 with Extra Trees (App. L.17), so most of the raw score is population structure the adversary already holds, and it becomes a practical risk mainly when the adversary also holds quasi-identifiers correlated with the sensitive column (App. F.2). An optional hardening step cuts attribute inference on the real-data core at -0.9% utility ($0.77 \to 0.52$, Table 21) and at a total -2.8 pp against the real context on the 14-tier ($0.59 \to 0.48$, Table 22). It adds noise at three generator points (App. F.1) while membership stays at MIA ≈ 0.55, and the DP-on-means layer is a preliminary robustness check rather than a full DP claim.

Ablating the generator on the 14-tier shows the 50:50 GMM/copula ensemble at -1.4 points, ahead of either component alone and of one Gaussian per class, while perturbing the real rows never matches its membership protection, and the closest setting costs -5.1 points (Table 56, App. L.11).

7 Conclusion

We introduced the Synthetic Data Shield (SDS), which replaces a TFM's in-context training set with a locally generated synthetic one to keep the real records private. Evaluated across 71 datasets and six in-context readers, the framework preserves accuracy while suppressing measured membership and near-record leakage, leaving attribute inference as the residual risk that tracks accuracy.

AI USE STATEMENT

Generative AI (Claude, Anthropic) assisted us with LaTeX editing and document formatting, with revisions for prose clarity, with checking citations against their primary sources, with organizing and tabulating the experimental results that our own scripts produced, and with reviewing the code of the released experiment and analysis scripts. It was not used to design the core method, to run experiments without human oversight, or to produce the theoretical claims, the proofs, or the numerical results reported in this paper. It was not used to generate data, to propose hypotheses, or to interpret results, and translation, dataset cleaning, and qualitative data analysis are not applicable to this work. Every reported number comes from the code released with this submission, executed on our own compute infrastructure and verified by the authors. We reviewed all AI-assisted text and code changes, and we take responsibility for the final content of this work, including any text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to reduce the privacy exposure created when tabular foundation models are consumed as hosted services. All experiments use publicly available benchmark datasets from the UCI and OpenML repositories or synthetic data generated with scikit-learn, and involve no human subjects, personally identifiable information, or proprietary data. SDS is an obfuscation layer rather than a formal privacy mechanism. It does not provide the per-record guarantee of differential privacy, and the paper reports the residual attribute-inference leakage openly (§6, App. F.2) together with explicit guidance on when not to use the method as-is (App. M). Practitioners handling regulated data should base deployment decisions on the audited leakage numbers rather than on the protection the method's name suggests.

REPRODUCIBILITY STATEMENT

The supplementary material includes the full pipeline, attack implementations, and evaluation code, along with the exact OpenML task IDs and the dataset-alias map (App. A). Algorithm 1 and App. D specify the generators in pseudocode, §6 and the appendix tables record every protocol setting (folds, seeds, α, K), App. C contains complete proofs of all theoretical results, and hardware details appear in App. F.

REFERENCES

Meenatchi Sundaram Muthu Selva Annamalai, Andrea Gadotti, and Luc Rocher. A linear reconstruction approach for attribute inference attacks against synthetic data. In *USENIX Security Symposium*, 2024.

Sergul Aydore, William Brown, Michael Kearns, Krishnaram Kenthapadi, Luca Melis, Aaron Roth, and Amaresh Ankit Siva. Differentially private query release through adaptive projection. In *International Conference on Machine Learning (ICML)*, 2021.

Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael Gomes Mantovani, Jan N. van Rijn, and Joaquin Vanschoren. OpenML benchmarking suites. *arXiv preprint arXiv:1708.03731*, 2019.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, 2006. ISBN 978-0-387-31073-2.

Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. The Johnson–Lindenstrauss transform itself preserves differential privacy. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 410–419. IEEE, 2012. doi: 10.1109/FOCS.2012.67. URL `https://arxiv.org/abs/1204.2136`. Conference version in FOCS 2012; preprint: arXiv:1204.2136.

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. In *International Conference on Learning Representations (ICLR)*, 2023.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Alycia N. Carey, Karuna Bhaila, Kennedy Edemacu, and Xintao Wu. DP-TabICL: In-context learning with differentially private tabular data. In *IEEE International Conference on Big Data (BigData)*, 2024. arXiv:2403.05681.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy (S&P)*, 2022.

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

Trevor F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman & Hall/CRC, Boca Raton, FL, 2nd edition, 2000.

Felix den Breejen, Sangmin Bae, Stephen Cha, and Se-Young Yun. Fine-tuned in-context learning transformers are excellent tabular data classifiers. *arXiv preprint arXiv:2405.13396*, 2024.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Stochastic Modelling and Applied Probability*. Springer, New York, 1996.

Dheeru Dua and Casey Graff. UCI machine learning repository. `http://archive.ics.uci.edu/ml`, 2017.

Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. Flocks of stochastic parrots: Differentially private prompt learning for large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–487, 2014. doi: 10.1561/0400000042. URL `https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf`.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, volume 3876 of *Lecture Notes in Computer Science*, pp. 265–284. Springer, 2006. doi: 10.1007/11681878_14. URL `https://link.springer.com/chapter/10.1007/11681878_14`.

Mei Ling Fang, Devendra Singh Dhami, and Kristian Kersting. DP-CTGAN: Differentially private medical data generation using CTGANs. In *International Conference on Artificial Intelligence in Medicine (AIME)*, 2022.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3–4):707–738, 2015.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning, 2017. Lipschitz properties of softmax established in Lemma 1.

Fengyu Gao, Ruida Zhou, Tianhao Wang, Cong Shen, and Jing Yang. Data-adaptive differentially private prompt synthesis for in-context learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Josh Gardner, Juan C. Perdomo, and Ludwig Schmidt. Large scale transfer learning for tabular data via language modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Oded Goldreich. *Foundations of Cryptography, Volume 2: Basic Applications*. Cambridge University Press, 2004.

Steven Golob, Sikha Pentyala, and Martine De Cock. SoK: Reconstruction attacks on synthetic tabular data (insights from winning the NIST CRC). *arXiv preprint arXiv:2606.08372*, 2026.

Google Research. TabFM: A zero-shot foundation model for tabular data, 2026. `https://research.google/blog/introducing-tabfm-a-zero-shot-foundation-model-for-tabular-data/`.

John C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. doi: 10.1007/BF02291478.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, et al. TabPFN-2.5: Advancing the state of the art in tabular foundation models. *arXiv preprint arXiv:2511.08667*, 2025.

Manbir S. Gulati and Paul F. Roysdon. TabMT: Generating tabular data with masked transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Meng Hao, Hongwei Li, Hanxiao Chen, Pengzhi Xing, Guowen Xu, and Tianwei Zhang. Iron: Private inference on transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Frederik Harder, Kamil Adamczewski, and Mijung Park. DP-MERF: Differentially private mean embeddings with random features for practical privacy-preserving data generation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. TabLLM: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=cp5PvcI6w8_.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6. URL https://www.nature.com/articles/s41586-024-08328-6.

Junyuan Hong, Jiachen T. Wang, Chenhui Zhang, Zhangheng Li, Bo Li, and Zhangyang Wang. DP-OPT: Make large language model your privacy-preserving prompt engineer. In *International Conference on Learning Representations (ICLR)*, 2024.

Florimond Houssiau, James Jordon, Samuel N. Cohen, Owen Daniel, Andrew Elliott, James Geddes, Callum Mole, Camila Rangel-Smith, and Lukasz Szpruch. TAPAS: a toolbox for adversarial privacy auditing of synthetic data. In *NeurIPS Workshop on Synthetic Data for Empowering ML Research*, 2022.

Alexia Jolicoeur-Martineau, Kilian Fatras, and Tal Kachman. Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

James Jordon, Jinsung Yoon, and Mihaela van der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations (ICLR)*, 2019.

James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Sherborne, et al. Synthetic data – what, why and how? *arXiv preprint arXiv:2205.03257*, 2022.

Myung Jun Kim, Léo Grinsztajn, and Gaël Varoquaux. CARTE: Pretraining and transfer for tabular learning. In *International Conference on Machine Learning (ICML)*, 2024.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17564–17579. PMLR, 2023. URL https://proceedings.mlr.press/v202/kotelnikov23a.html.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Brian G. Leroux. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20(3):1350–1360, 1992. doi: 10.1214/aos/1176348772. URL https://www.jstor.org/stable/2242000.

LimiX Team. LimiX: Unleashing structured-data modeling capability for generalist intelligence. *arXiv preprint arXiv:2509.03505*, 2025.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Yehuda Lindell and Benny Pinkas. Secure multiparty computation for privacy-preserving data mining. *Journal of Privacy and Confidentiality*, 1(1), 2009.

Tennison Liu, Zhaozhi Qian, Jeroen Berrevoets, and Mihaela van der Schaar. GOGGLE: Generative modelling for tabular data by learning relational structure. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Terrance Liu, Giuseppe Vietri, and Zhiwei Steven Wu. Iterative methods for private synthetic data: Unifying framework and new methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Hamidreza Kamkari, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L. Caterini, and Maksims Volkovs. TabDPT: Scaling tabular foundation models on real data. *arXiv preprint arXiv:2410.18164*, 2024.

Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Ryan McKenna, Daniel Sheldon, and Gerome Miklau. Graphical-model based estimation and inference for differential privacy. In *International Conference on Machine Learning (ICML)*, 2019.

Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Winning the NIST contest: A scalable and general approach to differentially private synthetic data. *Journal of Privacy and Confidentiality*, 11 (3), 2021. doi: 10.29012/jpc.778. URL https://arxiv.org/abs/2108.04978.

Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. AIM: An adaptive and iterative mechanism for differentially private synthetic data. *Proceedings of the VLDB Endowment*, 15(11):2599–2612, 2022. URL https://arxiv.org/abs/2201.12677.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2112.10510.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. *arXiv preprint arXiv:2502.05564*, 2025.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 20. Curran Associates, Inc., 2007. URL https://papers.nips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Donald B. Rubin. Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9(2): 461–468, 1993.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987. doi: 10.1007/BF01068419.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. doi: 10.1214/aos/1176344136. URL https://www.jstor.org/stable/2958889.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 3–18. IEEE, 2017.

Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, 8:229–231, 1959.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6048–6058, 2023.

Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data – anonymisation groundhog day. In *Proceedings of the 31st USENIX Security Symposium*, pp. 1451–1468. USENIX Association, 2022. URL https://www.usenix.org/conference/usenixsecurity22/presentation/stadler.

Xinyu Tang, Richard Shin, Huseyin A. Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. Privacy-preserving in-context learning with differentially private few-shot generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.11765.

Gianluca Truda. Generating tabular datasets under differential privacy. *arXiv preprint arXiv:2308.14784*, 2023.

Boris van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. DOMIAS: Membership inference attacks against synthetic data through overfitting detection. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

Veeravalli S. Varadarajan. On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics*, 19(1/2):23–26, 1958. URL https://www.jstor.org/stable/25048364.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2008. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9.

Dariush Wahdany, Matthew Jagielski, Jesse C. Cresswell, Adam Dziedzic, and Franziska Boenisch. TabPATE: Differentially private tabular in-context learning without public data. *arXiv preprint arXiv:2606.31474*, 2026.

Joshua Ward, Chi-Hua Wang, and Guang Cheng. Privacy auditing synthetic data release through local likelihood attacks. *arXiv preprint arXiv:2508.21146*, 2025.

Rui Wen, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against in-context learning. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2024.

Tong Wu, Ashwinee Panda, Jiachen T. Wang, and Prateek Mittal. Privacy-preserving in-context learning for large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2305.01639.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 7333–7343. Curran Associates, Inc., 2019. URL https://papers.nips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html.

Zexi Yao, Nataša Krčo, Georgi Ganev, and Yves-Alexandre de Montjoye. The DCR delusion: Measuring the privacy risk of synthetic data. In *European Symposium on Research in Computer Security (ESORICS)*, 2025.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium (CSF)*, 2018.

Jinsung Yoon, Lydia N. Drumright, and Mihaela van der Schaar. Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, 24(8):2378–2388, 2020.

Da Yu, Peter Kairouz, Sewoong Oh, and Zheng Xu. Privacy-preserving instructions for aligning large language models. In *International Conference on Machine Learning (ICML)*, 2024.

Hengrui Zhang, Jiani Zhang, Balasubramaniam Srinivasan, Zhengyuan Shen, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. PrivBayes: Private data release via Bayesian networks. *ACM Transactions on Database Systems*, 42(4):25:1–25:41, 2017. doi: 10.1145/3134428. Extended journal version; conference version in SIGMOD 2014.

Xiyuan Zhang, Danielle C. Maddix, Junming Yin, Nick Erickson, Abdul Fatir Ansari, Boran Han, Shuai Zhang, Leman Akoglu, Christos Faloutsos, Michael W. Mahoney, Yuyang Wang, Cuixiong Hu, Huzefa Rangwala, George Karypis, and Bing Wang. Mitra: Mixed synthetic priors for enhancing tabular foundation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y. Chen. CTAB-GAN: Effective table data synthesizing. In *Asian Conference on Machine Learning (ACML)*, 2021.

Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y. Chen. CTAB-GAN+: Enhancing tabular data synthesis. *Frontiers in Big Data*, 6, 2023.

Bingzhao Zhu, Xingjian Shi, Nick Erickson, Mu Li, George Karypis, and Mahsa Shoaran. XTab: Cross-table pretraining for tabular transformers. In *International Conference on Machine Learning (ICML)*, 2023.

Chaoyi Zhu, Jiayi Tang, Juan F. Pérez, Marten van Dijk, and Lydia Y. Chen. DP-TLDM: Differentially private tabular latent diffusion model. *arXiv preprint arXiv:2403.07842*, 2024.

A Extended Related Work

This section expands §3 with the full landscape.

In-context tabular foundation models. The models SDS evaluates (TabPFN and TabPFN-2.5 (Hollmann et al., 2023; 2025; Grinsztajn et al., 2025), TabICL and TabICLv2 (Qu et al., 2025; 2026), LimiX (LimiX Team, 2025), Google's TabFM (Google Research, 2026), and TabuLa-8B (Gardner et al., 2024)) are in-context readers of a labeled table. All but TabuLa-8B descend from prior-data fitted networks (Müller et al., 2022), which meta-train a transformer on synthetic priors to approximate the Bayesian posterior in one forward pass, while TabuLa-8B is a large language model fine-tuned on serialized tables. The wider family includes Amazon's Mitra (Zhang et al., 2025) and TabDPT (Ma et al., 2024), which, unlike the purely synthetic-prior models, are pretrained with

retrieval on real tables. It also includes cross-table and transfer approaches such as CARTE (Kim et al., 2024), XTab (Zhu et al., 2023), and the fine-tuned TabForestPFN (den Breejen et al., 2024), and LLM-serialization methods such as TabLLM (Hegselmann et al., 2023) that, like TabuLa-8B, cast rows as text. In every variant, the labeled context is transmitted at inference time, which is precisely the exposure SDS removes. SDS is model-agnostic and treats each of these as a black-box consumer of its synthetic context.

Tabular synthesizers and differentially private data release. Non-private deep synthesizers span GANs and autoencoders (CTGAN/TVAE (Xu et al., 2019), CTAB-GAN and CTAB-GAN+ (Zhao et al., 2021; 2023)), diffusion and score-based models (TabDDPM (Kotelnikov et al., 2023), TabSyn (Zhang et al., 2024), ForestDiffusion (Jolicoeur-Martineau et al., 2024)), masked-transformer (TabMT (Gulati & Roysdon, 2023)) and LLM (GReaT (Borisov et al., 2023)) generators, and relational models (GOGGLE (Liu et al., 2023)). All can memorize training rows (Somepalli et al., 2023; Yoon et al., 2020). Differentially private release covers graphical-model / select-measure-generate methods (PrivBayes (Zhang et al., 2017), MST and the NIST-winning approach (McKenna et al., 2021), Private-PGM (McKenna et al., 2019), AIM (McKenna et al., 2022)), iterative and query-release methods (GEM (Liu et al., 2021), RAP (Aydore et al., 2021)), and DP deep generators (PATE-GAN (Jordon et al., 2019), DP-CTGAN (Fang et al., 2022), DP-MERF (Harder et al., 2021), DP diffusion (Truda, 2023), and the DP latent-diffusion DP-TLDM (Zhu et al., 2024)). In the small-context regime, these either discretize away the fine-grained signal TabPFN needs or require far more rows per class than are available. SDS instead keeps the release non-private but accuracy-faithful and audits the residual leakage directly.

Privacy-preserving in-context learning. Beyond the prompt-level DP methods for language models cited in the body (Tang et al., 2024; Wu et al., 2024; Duan et al., 2023; Hong et al., 2024; Yu et al., 2024; Gao et al., 2025), the two closest works operate on *tabular* in-context learning: DP-TabICL (Carey et al., 2024) privatizes records or group statistics before serialization, and TabPATE (Wahdany et al., 2026) runs a PATE-style teacher aggregation without public data. Both accept a differential-privacy utility cost. SDS is DP-free and instead substitutes a synthetic context. We do not report a matched TabPATE comparison: it has no public implementation at the time of writing, and its teacher-partitioning and query-generation steps would have to be reconstructed from the paper, which risks understating it rather than comparing fairly. Membership-inference attacks against in-context learning (Wen et al., 2024) and against machine-learning models generally (Shokri et al., 2017; Yeom et al., 2018; Carlini et al., 2022) show that the context is worth protecting.

Synthetic-data privacy auditing. Our four-axis audit builds on the synthetic-data privacy literature: the "Groundhog Day" membership/attribute framework (Stadler et al., 2022), density-ratio membership inference (DOMIAS (van Breugel et al., 2023)), adversarial auditing toolkits (TAPAS (Houssiau et al., 2022)), linear-reconstruction attribute-inference attacks (Annamalai et al., 2024), and critiques of distance-to-closest-record heuristics (Yao et al., 2025). Stronger attacks continue to appear: no-box local-likelihood-ratio membership attacks that target local overfitting (Ward et al., 2025) and joint-structure reconstruction that couples per-feature posteriors on a dependency graph (Golob et al., 2026). We reuse these attack families but retarget them from "does a model trained on synthetic data generalize" to "does a synthetic *context* preserve a pre-trained model's predictions while hiding the real rows." We run the no-box local-likelihood-ratio attack (Gen-LRA) directly against SDS releases (App. F.19) and flag the joint-reconstruction attack as the remaining priority (App. M).

Dataset provenance. The evaluation spans 71 distinct datasets forming 106 dataset–protocol pairs across five suites, all drawn from public sources. (1) The *core controlled* six are Iris, Wine, and BreastCancer from the UCI repository (Dua & Graff, 2017) plus Multi3_Med, Multi5_Med, and HighDim_50 generated with scikit-learn's `make_classification` (Pedregosa et al., 2011). They are named in Tables 7 and 10. (2) The *repeated numeric headline* tier of 14 datasets (Table 11) is a UCI/OpenML mixture (Dua & Graff, 2017; Vanschoren et al., 2014) following the numeric-benchmark lineage of Grinsztajn et al. (2022). (3) The *mixed-type* tier of 15 datasets (Table 13) is from OpenML-CC18 (Bischl et al., 2019) accessed via OpenML (Vanschoren et al., 2014). (4) The *TabZilla-hard* tier of 31 datasets is the 36-dataset TabZilla benchmark (McElfresh et al., 2023) minus five that fail our class-count/size/loading filters (audiology, lymph, colic, albert, guillermo, App. M). (5) The *breadth* tier of 40 datasets is a wider OpenML-CC18/TabZilla/UCI mixture (Bischl et al.,

2019; McElfresh et al., 2023; Dua & Graff, 2017). The exact OpenML task IDs and the dataset-alias map (e.g. `kr-vs-kp`=Chess-KR, `car-eval`=Car, `german-credit`=Credit-g) are shipped in the supplementary code. Column counts d differ by a few across tables for the same dataset where a loader drops constant or identifier columns, and N differs where a protocol caps the row count.

B PRELIMINARIES

Throughout the appendix, p_{real} denotes the real data distribution written $p^\star$ in the body, and $\hat{\mu}_N^{\text{real}}$ denotes its empirical measure $\hat{\mu}_N$.

B.1 GAUSSIAN MIXTURE MODELS AND BIC SELECTION

A k-component GMM over $\mathbb{R}^d$ is $p_{\text{GMM}}(x) = \sum_{j=1}^{k} \pi_j \mathcal{N}(x; \mu_j, \Sigma_j)$. We fit class-conditional GMMs via Expectation-Maximization (EM) with BIC selection (Schwarz, 1978): $\text{BIC}(k) = -2 \log L_k + n_p(k) \log N_c$, choosing $\hat{k} = \arg\min_k \text{BIC}(k)$.

B.2 GAUSSIAN COPULA AND SKLAR'S THEOREM

Definition 1 (Copula, Sklar 1959)**.** A copula $C : [0,1]^d \to [0,1]$ is a joint CDF with uniform marginals. Sklar's theorem: every multivariate CDF F satisfies $F(x) = C(F_1(x_1), \ldots, F_d(x_d))$ for a unique C (continuous marginals).

The Gaussian copula parametrized by Σ is $C_\Sigma(u) = \Phi_d(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d); \Sigma)$.

B.3 WASSERSTEIN DISTANCE, KULLBACK–LEIBLER DIVERGENCE, AND JENSEN–SHANNON DIVERGENCE

The Kullback–Leibler divergence between probability distributions p, q on a common measurable space is $\text{KL}(p\|q) := \int p \log(p/q)\, d\mu_0$ (with μ_0 a dominating measure, and we write $\int p \log(p/q)$ for short). The 1-Wasserstein distance and its Kantorovich–Rubinstein dual are

$$\mathcal{W}_1(\mu, \nu) = \inf_\gamma \int \|x - y\| \; d\gamma(x, y) = \sup_{\|f\|_L \le 1} \Big| \int f \, d\mu - \int f \, d\nu \Big|,$$

where $\|f\|_L := \sup_{x \neq y} |f(x) - f(y)| / \|x - y\|$ is the Lipschitz seminorm. The Jensen–Shannon divergence is $\text{JSD}(p\|q) = \frac{1}{2}\text{KL}(p\|m) + \frac{1}{2}\text{KL}(q\|m)$ with $m = \frac{1}{2}(p + q)$.

C PROOFS

C.1 SCOPE OF THE ANALYSIS

SDS's components receive three levels of support. *Covered theoretically*: Theorems 1–3 identify the operative quantities for utility, membership, and attribute leakage, with Theorem 2 stated for the full release law, and Proposition 1 characterizes geometry decoupling under i.i.d. sampling. *Conditionally analyzed*: the numeric branch is stable under explicit conditions: for the GMM, an unchanged BIC-selected component count and convergence to the same basin of the expectation-maximization (EM) fit (Proposition 7), and for the copula, clipping with covariance regularization (Remark 6). *Evaluated empirically*: candidate selection, the DCR filter, and the categorical pipeline are supported by audits of the final released contexts (Table 40, App. J.4, App. F.25). We do not claim an end-to-end certificate for the full selected-release pipeline. The headline runs sample a single context (K=1), so the audited release is the i.i.d. release law the theorems address.

Remark 1 (From bounded loss to the reported R^2). Theorem 3 requires $\ell \in [0,1]$, while the attribute-inference audit reports Ridge R^2 (squared loss, unbounded). On the standardized, percentile-clipped features used throughout, define the truncated loss $\ell_B(y, \hat{y}) = \min((y - \hat{y})^2, B)/B \in [0,1]$ for a constant B. Applying the theorem to ℓ_B gives $R_P(h_G) \le \inf_h R_P(h) + 2B \cdot \text{TV}(p^\star, p_G)$ for the clipped risk, so the statement covers the truncated-loss analogue of R^2. It does not extend to the uncapped squared loss we report, because Ridge predictions are unbounded and the tail beyond B is

uncontrolled. We report uncapped R^2 per column for comparability with Stadler et al. (2022) and treat it as an empirical quantity. Where per-column values are averaged (Tables 44 and 62) each is first clipped at -1 so a single degenerate fit cannot dominate the mean.

C.2 Proof of Theorem 3

Proof. Write $R_P(h) = \mathbb{E}_{p^\star}\big[\ell(h)\big]$ and $R_G(h) = \mathbb{E}_{p_G}\big[\ell(h)\big]$ for the risks of a predictor $h \in \mathcal{H}$ under the real and generator distributions. Because $\ell \in [0,1]$, the variational characterisation of total variation,

$$\mathrm{TV}(p^\star, p_G) \;=\; \sup_{0\le f\le 1}\Big|\mathbb{E}_{p^\star}[f] - \mathbb{E}_{p_G}[f]\Big|,$$

applies to $f = \ell(h)$ for every h, so the two risks differ by at most the total variation distance uniformly over the class,

$$\sup_{h\in\mathcal{H}}\big|R_P(h) - R_G(h)\big| \;\le\; \mathrm{TV}(p^\star, p_G). \tag{6}$$

Let $h^\star \in \arg\min_{h\in\mathcal{H}} R_P(h)$ be a best real-data predictor in the class, and recall $h_G \in \arg\min_{h\in\mathcal{H}} R_G(h)$. Applying (6) to h_G, then optimality of h_G under R_G, then (6) again to $h^\star$,

$$R_P(h_G) \;\le\; R_G(h_G) + \mathrm{TV}(p^\star, p_G) \;\le\; R_G(h^\star) + \mathrm{TV}(p^\star, p_G) \;\le\; R_P(h^\star) + 2\,\mathrm{TV}(p^\star, p_G),$$

which is the claim, since $R_P(h^\star) = \inf_{h\in\mathcal{H}} R_P(h)$. □

The bound is one-sided: it limits how much worse a generator-trained attribute predictor can be than the best real-data predictor in the same class. It does not assert that a small TV forces attribute inference to succeed, so it supports the statement that a high-fidelity generator *can* transfer attribute-predictive structure, not that every utility-preserving generator must (Remark 1 covers the passage from this bounded loss to the unbounded squared loss the audit reports).

C.3 Proof of Theorem 1

Proof. By context-Lipschitzness,

$$\|f_\theta(x_{\text{test}}; \hat{\mu}_n) - f_\theta(x_{\text{test}}; \hat{\mu}_N^{\text{real}})\|_\infty \le L_f\, \mathcal{W}_1(\hat{\mu}_n, \hat{\mu}_N^{\text{real}}).$$

By the triangle inequality on $\mathcal{W}_1$,

$$\mathcal{W}_1(\hat{\mu}_n, \hat{\mu}_N^{\text{real}}) \le \mathcal{W}_1(\hat{\mu}_n, p_G) + \mathcal{W}_1(p_G, p_{\text{real}}) + \mathcal{W}_1(p_{\text{real}}, \hat{\mu}_N^{\text{real}}).$$

Taking expectations gives the bound with $R_n = \mathbb{E}[\mathcal{W}_1(\hat{\mu}_n, p_G)] = O(n^{-1/d})$ and $R_N = \mathbb{E}[\mathcal{W}_1(p_{\text{real}}, \hat{\mu}_N^{\text{real}})] = O(N^{-1/d})$ (Fournier & Guillin, 2015). □

C.4 Context-Lipschitz: Assumption, Proposition, and Proof

The body's Theorem 1 is stated under a context-Lipschitz assumption on the pre-trained TabPFN map. Assumption 1 below is stated for the expected predictor over resampled contexts of a fixed size, whereas Theorem 1 compares two actual contexts of different sizes. Propositions 2 and 4 are the separate sufficient conditions under which the actual context map satisfies the assumption.

Assumption 1 (Context-Lipschitz, general case)**.** Fix f_θ, $x_{\text{test}} \in \mathbb{R}^d$, and a compact set $\mathcal{K} \subset \mathbb{R}^d$. For each context size $n \ge 1$, define

$$f_{\theta,n}(x^*; \mu) := \mathbb{E}_{(X,y)\sim\mu^n}[f_\theta(x^*; X, y)], \qquad \mu \in \mathcal{P}(\mathcal{K} \times [C]).$$

The assumption is that for the context sizes used in this paper there exists $L_f \in (0,\infty)$ such that $\|f_{\theta,n}(x_{\text{test}}; \mu) - f_{\theta,n}(x_{\text{test}}; \nu)\|_\infty \le L_f \cdot \mathcal{W}_1(\mu, \nu)$ for all $\mu, \nu \in \mathcal{P}(\mathcal{K} \times [C])$.

Proposition 2 (Context-Lipschitz for equal-size empirical measures)**.** *Let f_θ be permutation-invariant with $\tilde{L} \in (0,\infty)$ a Lipschitz constant for its output under the entry-wise ℓ_1 norm $\|\cdot\|_1$, and $h : \mathbb{R}^d \times [C] \to \mathbb{R}^p$ a token-embedding map with constant $B_{\text{emb}} \in (0,\infty)$. For equal-size empirical measures $\hat{\mu}_n, \hat{\nu}_n$ of size n, $\|f_\theta(x_{\text{test}}; \hat{\mu}_n) - f_\theta(x_{\text{test}}; \hat{\nu}_n)\|_\infty \le \tilde{L} \cdot B_{\text{emb}} \cdot n \cdot \mathcal{W}_1(\hat{\mu}_n, \hat{\nu}_n)$.*

Proof. For uniform empirical measures of size n, the optimal $\mathcal{W}_1$ coupling is a bijection π^*. Relabel so that $z_i' = \pi^*(z_i)$. Then $\mathcal{W}_1 = \frac{1}{n}\sum_i \rho(z_i, z_i')$. By permutation invariance, evaluate both predictions using this ordering. Token matrices H, H' satisfy $\|H - H'\|_1 = \sum_i \|h(z_i) - h(z_i')\|_1 \le B_{\text{emb}} \sum_i \rho(z_i, z_i') = B_{\text{emb}} n \mathcal{W}_1$. Apply $\tilde{L}$-Lipschitz. □

C.5 Token-Level Stability of the Pre-trained Transformer

Proposition 3 (Token-level stability). *Let $f_\theta^{\mathrm{tok}} : \mathbb{R}^{(m+1)\times p} \to \Delta^{C-1}$ denote the pre-trained TabPFN forward map acting on a token matrix $H \in \mathbb{R}^{(m+1)\times p}$, with $\|\cdot\|_F$ the Frobenius norm. On any compact reached region with bounded attention logits, Lipschitz activations, and layer-norm input variance bounded below by $\eta \in (0,\infty)$, there is a finite Lipschitz constant $L_{\mathrm{tok}} \in (0,\infty)$ depending only on the network weights, depth, and norm bounds, such that $\|f_\theta^{\mathrm{tok}}(H) - f_\theta^{\mathrm{tok}}(H')\|_\infty \le L_{\mathrm{tok}} \cdot \|H - H'\|_F$ for any equal-shape H, H' in the region.*

Proof. Each transformer sublayer is locally Lipschitz on the regular reached region: linear projections via operator norm $\|W\|_{1\to 1}$, ReLU as 1-Lipschitz and GELU under a finite derivative bound on bounded domains, and softmax with bounded logits is Lipschitz under $\|\sigma(a) - \sigma(a')\|_1 \le \|a - a'\|_1$ (Gao & Pavel, 2017). Layer normalization is Lipschitz when its input variance is bounded below by η. Residual connections preserve Lipschitzness by addition. Feed-forward blocks are finite compositions of the above. Composing the finitely many L layers yields a finite product $L_{\mathrm{tok}} = M^L$, where $M \in (0,\infty)$ is the per-layer Lipschitz factor implied by the weights and norm bounds. Apply this to two equal-shape token matrices H, H' in the reached region. □

The product bound M^L is vacuously large for TabPFN ($L = 24$ for the TabPFN-2.5 reader). The empirical diagnostic in App. E supports bounded context sensitivity (median ratio 20.3, $p_{95} = 112.8$) but is not a direct estimate of L_{tok}. Proposition 3 is *local* (it holds on the compact reached region) and *architectural*. It does not assume TabPFN is a Wasserstein-Lipschitz functional of the data distribution. Theorem 1 bounds $\|H_{\mathrm{synth}} - H_{\mathrm{real}}\|_F$ via $\mathcal{W}_1(p_{\mathrm{real}}, p_G)$ and is therefore a sufficient distributional condition rather than an architectural one.

C.6 Oversampling Extension and General-Measure Lipschitz

The body's Wasserstein utility theorem (Theorem 1) and Proposition 2 cover $\alpha{=}1$ via a bijection between equal-size empirical measures. For $\alpha{>}1$ the bijection is unavailable, and we use a mean-aggregation abstraction.

Proposition 4 (Context-Lipschitz for general probability measures). *Let f_θ be a predictor whose output for query x_{test} given context measure μ can be written as $f_\theta(x_{\mathrm{test}}; \mu) = g_\theta\big(x_{\mathrm{test}}, \mathbb{E}_{z\sim\mu}[\varphi_\theta(x_{\mathrm{test}}, z)]\big)$, where φ_θ is B_φ-Lipschitz in z and g_θ is L_g-Lipschitz in its second argument. Then for any two Borel probability measures μ, ν on $\mathcal{K}\times[C]$: $\|f_\theta(x_{\mathrm{test}}; \mu) - f_\theta(x_{\mathrm{test}}; \nu)\|_\infty \le L_g B_\varphi \cdot \mathcal{W}_1(\mu, \nu)$.*

Remark 2 (Scope of the decomposition). Proposition 4 idealises TabPFN as $g_\theta \circ \mathbb{E}_\mu[\varphi_\theta]$. In multi-layer self-attention, token interactions are data-dependent and nonlinear. Proposition 2 is exact for $\alpha{=}1$. The two propositions cover separate regimes: at $\alpha{=}1$ under permutation invariance and a Lipschitz token map, and under the mean-aggregation abstraction at $\alpha{>}1$. We treat the result as a sufficient-condition heuristic.

Corollary 1 (Direct bound for oversampled protocols ($\alpha \ge 1$)). *For $\alpha \ge 1$ and $\hat{\mu}_{\alpha N}$ the empirical measure of αN i.i.d. synthetic draws from p_G: (i) $\mathbb{E}[\|f_\theta(x_{\mathrm{test}}; \hat{\mu}_{\alpha N}) - f_\theta(x_{\mathrm{test}}; \hat{\mu}_N^{\mathrm{real}})\|_\infty] \le L_g B_\varphi[\mathcal{W}_1(p_G, p_{\mathrm{real}}) + R_{\alpha N} + R_N]$, $R_m{=}O(m^{-1/d})$ (Fournier & Guillin, 2015), (ii) the synthetic finite-sample term has the smaller leading order for $\alpha{>}1$, since $(\alpha N)^{-1/d} < N^{-1/d}$, though the rates alone do not fix the constants, (iii) empirically $\bar{\Delta}(\alpha{=}2){=}-1.4\%$ vs. $\bar{\Delta}(\alpha{=}1){=}-1.8\%$ on the 14-tier (Table 29), with 13/14 datasets showing equal or smaller absolute gap at $\alpha{=}2$.*

Remark 3 (Effective Lipschitz constant). Composing $L{=}24$ transformer layers gives $\tilde{L} \le M^{24}$ (vacuously large). The empirical diagnostic (App. E) shows the effective constant is moderate (median 20.3, $p_{95}{=}112.8$), standard for the gap between per-layer product bounds and end-to-end sensitivity.

C.7 Proof of Proposition 4

Proof. Under the mean aggregation abstraction stated in Proposition 4, the context enters through

$$\mathbb{E}_{z\sim\mu}[\varphi_\theta(x_{\mathrm{test}}, z)].$$

Thus the predictor is modeled as

$$f_\theta(x_{\text{test}}; \mu) = g_\theta(x_{\text{test}}, \mathbb{E}_{z\sim\mu}[\varphi_\theta(x_{\text{test}}, z)]) .$$

For two measures μ, ν (of potentially different support sizes):

$$\begin{aligned}\|\mathbb{E}_\mu[\varphi] - \mathbb{E}_\nu[\varphi]\| &= \left\| \int \varphi(x_{\text{test}}, z)\, d\mu(z) - \int \varphi(x_{\text{test}}, z)\, d\nu(z) \right\| \\ &\le B_\varphi \cdot \mathcal{W}_1(\mu, \nu)\end{aligned}$$

by Kantorovich–Rubinstein duality (since $z \mapsto \varphi(x_{\text{test}}, z)$ is B_φ-Lipschitz). This holds for any Borel probability measures μ, ν on $\mathcal{K} \times [C]$, with no requirement that they have the same support size. Composing with the L_g-Lipschitz readout: $\|f_\theta(\mu) - f_\theta(\nu)\|_\infty \le L_g \|\mathbb{E}_\mu[\varphi] - \mathbb{E}_\nu[\varphi]\| \le L_g B_\varphi \cdot \mathcal{W}_1(\mu, \nu)$. □

Under this abstraction the constant $L_g B_\varphi$ replaces $\tilde{L} B_{\text{emb}} n$ from the equal-size case (Proposition 2) and no longer depends on context size.

C.8 Proof of Corollary 1

Proof. **(i)** Apply Proposition 4 with $\mu = \hat{\mu}_{\alpha N}$ and $\nu = \hat{\mu}_N^{\text{real}}$ (empirical measures of sizes αN and N respectively, no bijection required). Triangle inequality on $\mathcal{W}_1$: $\mathcal{W}_1(\hat{\mu}_{\alpha N}, \hat{\mu}_N^{\text{real}}) \le \mathcal{W}_1(\hat{\mu}_{\alpha N}, p_G) + \mathcal{W}_1(p_G, p_{\text{real}}) + \mathcal{W}_1(p_{\text{real}}, \hat{\mu}_N^{\text{real}})$. Take expectations. The finite-sample rates give $R_{\alpha N} + R_N$ where $R_m = O(m^{-1/d})$.

(ii) $R_{\alpha N} = O((\alpha N)^{-1/d})$ and $R_N = O(N^{-1/d})$, so for $\alpha > 1$ the synthetic term has the smaller leading order. This compares rates, not the constants.

(iii) Empirically (Table 29), α=2 improves the mean accuracy gap relative to α=1 ($\bar{\Delta}(\alpha{=}2) = -1.4\%$ vs. $\bar{\Delta}(\alpha{=}1) = -1.8\%$). The improvement is not pointwise. Ionosphere is slightly worse at α=2, whereas 13 of 14 datasets show an equal or smaller absolute gap. □

C.9 Distributional-Property Comparison with K-RFF

Table 4 contrasts SDS with the K-RFF release (App. F.4) on the three structural properties controlled by the theory of §5: geometry decoupling, prediction stability, and the requirement of a shared secret key between data owner and prediction service. SDS achieves geometry decoupling by Proposition 1 and prediction stability under Assumption 1 together with Proposition 2. K-RFF is designed to preserve accuracy through a distance-preserving feature map, which is also what leaves pairwise distances recoverable (R^2=0.735, Table 10), and it additionally requires the prediction service to hold the projection key.

Table 4: Distributional-property comparison: SDS vs. K-RFF.

Property	**SDS**	**K-RFF**
Geometry decoupling	Prop. 1, generator level only	Distances preserved by design
Prediction-stability coupling	Asm. 1; Prop. 2	None
Requires shared secret key	No	Yes

C.10 Ensemble KL Bound

Proposition 5 (Ensemble KL Bound)**.** $\mathrm{KL}(p_{\text{real}}(\cdot \mid c) \| p_{\text{ens}}(\cdot \mid c)) \le \frac{1}{2}\mathrm{KL}(p_{\text{real}}(\cdot \mid c) \| p_{\text{GMM}}(\cdot \mid c)) + \frac{1}{2}\mathrm{KL}(p_{\text{real}}(\cdot \mid c) \| p_{\text{Cop}}(\cdot \mid c))$.

Proof. Since $-\log$ is convex, Jensen's inequality gives $-\log(\frac{1}{2}q_1 + \frac{1}{2}q_2) \le -\frac{1}{2}\log q_1 - \frac{1}{2}\log q_2$. Multiply by p_{real} and integrate: $\mathrm{KL}(p_{\text{real}} \| p_{\text{ens}}) \le \frac{1}{2}\mathrm{KL}(p_{\text{real}} \| p_{\text{GMM}}) + \frac{1}{2}\mathrm{KL}(p_{\text{real}} \| p_{\text{Cop}})$. □

C.11 Proof of Proposition 1

Proof. **(a)** Varadarajan's theorem (Varadarajan, 1958) gives $\hat{\mu}_n \to p_G^X$ weakly a.s., and product convergence + continuity of $(x, y) \mapsto \|x - y\|$ yields convergence of pairwise-distance CDFs. **(b)** Continuous mapping theorem. **(c)** A k-component GMM with $k < N$ has population distances that generically differ from the discrete distances of N points. □

C.12 Discrimination–TV Identity

Proposition 6 (Discrimination–TV identity)**.** *Under equal priors, the optimal balanced accuracy of any classifier discriminating p_{real} from p_G is*

$$\text{Acc}^*_{\text{bal}} \;=\; \tfrac{1}{2}\big(1 + \text{TV}(p_{\text{real}}, p_G)\big) \;\le\; \tfrac{1}{2} + \sqrt{\text{JSD}(p_{\text{real}} \| p_G)/2},$$

where JSD *is the Jensen–Shannon divergence. This is an identity relating discrimination accuracy to* $\text{TV}(p_{\text{real}}, p_G)$*. It does not assert* TV *is small, so a high discriminator AUC (e.g., Segment, App. F) is fully consistent with near-chance membership advantage on the same release, since discrimination and record-level identification are governed by different quantities.*

Proof. Bayes error $= \frac{1}{2}(1 - \text{TV})$ (Devroye et al., 1996), and $\text{TV} \le \sqrt{2\text{JSD}}$ (Lin, 1991). Combining gives $\text{Acc}^*_{\text{bal}} = \frac{1}{2}(1 + \text{TV}) \le \frac{1}{2} + \sqrt{\text{JSD}/2}$. □

Remark 4 (Extension to non-Lipschitz adversaries). For adversaries with Lipschitz constant $L > 1$, the bound becomes $L \cdot \mathcal{W}_1(p^X_{\text{real}}, p^X_G)$, covering linear classifiers and shallow networks. Deep networks are not covered, motivating the neural discriminator as empirical complement.

C.13 Proof of Theorem 2

Proof. Fix a neighbouring pair $\mathcal{D} \simeq \mathcal{D}'$ and let the adversary A observe a single release R drawn from $Q_{\mathcal{D}}$ or $Q_{\mathcal{D}'}$ with equal prior probability $\frac{1}{2}$. Deciding which dataset produced R is a binary hypothesis test between $Q_{\mathcal{D}}$ and $Q_{\mathcal{D}'}$ under equal priors, so the smallest achievable error probability is the Bayes error

$$\text{err}^\star = \tfrac{1}{2}\big(1 - \text{TV}(Q_{\mathcal{D}}, Q_{\mathcal{D}'})\big),$$

and the largest achievable accuracy is $1 - \text{err}^\star = \frac{1}{2}\big(1 + \text{TV}(Q_{\mathcal{D}}, Q_{\mathcal{D}'})\big)$ (Devroye et al., 1996). Since $\text{Adv}(A)$ is the accuracy of A above the chance level $\frac{1}{2}$,

$$\text{Adv}(A) \;\le\; \tfrac{1}{2}\,\text{TV}(Q_{\mathcal{D}}, Q_{\mathcal{D}'}).$$

Pinsker's inequality gives $\text{TV}(Q_{\mathcal{D}}, Q_{\mathcal{D}'}) \le \sqrt{\frac{1}{2}\text{KL}(Q_{\mathcal{D}} \| Q_{\mathcal{D}'})}$, and by the definition of Γ_∞ as a supremum over neighbouring pairs of the larger of the two divergences, $\text{KL}(Q_{\mathcal{D}} \| Q_{\mathcal{D}'}) \le \Gamma_\infty$ for every such pair. Hence

$$\text{Adv}(A) \;\le\; \tfrac{1}{2}\sqrt{\Gamma_\infty/2} \;=\; \sqrt{\Gamma_\infty/8}.$$

The bound holds for every neighbouring pair and every balanced adversary, and the right-hand side does not depend on the pair, so it holds uniformly. □

Corollary 2 restates this bound in the form used by the numerical diagnostics of App. L.8. The quantity Γ_∞ is a supremum over neighbours of the *full* release law, while the empirical $\widehat{\Gamma}$ of Table 5 is a sampled maximum over perturbations of the GMM component only, which is why we treat the resulting numbers as a diagnostic rather than a certified ceiling.

C.14 Pinsker Form of Theorem 2

Corollary 2 (Pinsker membership bound)**.** *Under the conditions of Theorem 2, the advantage of any balanced replace-one membership adversary is at most $\sqrt{\Gamma_\infty/8}$.*

Proof. Pinsker's inequality gives $\text{TV}(Q_{\mathcal{D}}, Q_{\mathcal{D}'}) \le \sqrt{\text{KL}(Q_{\mathcal{D}} \| Q_{\mathcal{D}'})/2} \le \sqrt{\Gamma_\infty/2}$. The advantage of a balanced replace-one distinguisher equals $\frac{1}{2}\text{TV}$ (Devroye et al., 1996), giving the claim. □

On Wine and BreastCancer the cumulative $\widehat{\Gamma}$ in Table 5 pushes $\sqrt{\widehat{\Gamma}/8}$ beyond 0.5, while the empirically audited NN-MIA / QAPD advantages stay at 0.03–0.06.

C.15 Conditional Local Stability of the GMM/Copula Stack

Theorem 2 is mechanism-agnostic: it identifies the replace-one KL of the actual release law as the governing quantity. We do *not* treat SDS as a strongly-convex regularized ERM (EM is non-convex and BIC is a discrete operator). The proposition is explicitly conditional, and Theorem 2 carries the unconditional claim. Let $\mathcal{D} \simeq \mathcal{D}'$ be neighboring datasets. Per class $c \in [C]$, assume

(i) **bounded data:** $\|x\|_2 \le R$ for some $R \in (0, \infty)$ after preprocessing
(ii) **covariance floor:** $\lambda_{\min}(\Sigma_{cj}) \ge \sigma_{\min}^2$ for some $\sigma_{\min} \in (0, \infty)$ and every fitted component covariance Σ_{cj} (the covariance of GMM component j in class c)
(iii) **BIC margin:** the BIC-selected $k_c^\star$ has margin $\ge 2B_{\mathrm{BIC}}$ over its runner-up, with $B_{\mathrm{BIC}} \in (0, \infty)$, and replacing one record changes every candidate score by at most B_{BIC}
(iv) **local EM stability:** conditional on the same $k_c^\star$, EM converges to a local optimum $\hat{\theta}_c(\mathcal{D})$ whose negative log-likelihood is μ_{sc}-strongly convex in a neighborhood of the optimum, for some $\mu_{\mathrm{sc}} \in (0, \infty)$, with per-example score-norm bound $B_{\mathrm{sc}} \in (0, \infty)$.

Proposition 7 (Conditional local stability)**.** *Under (i)–(iv), $\|\hat{\theta}_c(\mathcal{D}) - \hat{\theta}_c(\mathcal{D}')\|_2 \le 2B_{\mathrm{sc}}/(\mu_{\mathrm{sc}} N_c)$. If the class-conditional family is locally KL-smooth in the sense $\mathrm{KL}(G_{\theta_c} \| G_{\theta_c'}) \le \frac{L_{\mathrm{KL}}}{2}\|\theta_c - \theta_c'\|^2$ for some $L_{\mathrm{KL}} \in (0, \infty)$, where G_{θ_c} denotes the GMM parameterised by θ_c, then for m i.i.d. synthetic rows*

$$\sup_A \mathrm{Adv}(A) \;\le\; \frac{B_{\mathrm{sc}}}{2\mu_{\mathrm{sc}} N_{\min}} \sqrt{m \cdot L_{\mathrm{KL}}}, \qquad N_{\min} := \min_c N_c.$$

For $m = \alpha N$ with bounded class proportions, this scales as $O(\sqrt{\alpha/N})$.

Proof. The BIC margin (iii) ensures $\mathcal{D}$ and $\mathcal{D}'$ select the same $k_c^\star$, ruling out a discrete model-selection jump. Conditional on the same model and EM basin, the fitted parameter is a local M-estimator of $F_{\mathcal{D}}(\theta) := \frac{1}{N_c}\sum_{z_i \in \mathcal{D}_c} \ell(\theta; z_i)$. Strong convexity and first-order optimality give $\mu_{\mathrm{sc}} \cdot \|\hat{\theta}_c(\mathcal{D}) - \hat{\theta}_c(\mathcal{D}')\| \le \|\nabla F_{\mathcal{D}}(\hat{\theta}_c(\mathcal{D}')) - \nabla F_{\mathcal{D}'}(\hat{\theta}_c(\mathcal{D}'))\|$. Since $\mathcal{D}, \mathcal{D}'$ differ in one record and per-example score norms are bounded by B_{sc}, the right-hand side is at most $2B_{\mathrm{sc}}/N_c$. Apply KL-smoothness, the additivity of KL over the m i.i.d. synthetic rows, and Theorem 2. □

Remark 5 (Discrete jumps). Proposition 7 is intentionally conditional. If the BIC gap is small, replacing one record may change the selected component count. If EM converges to a different basin, local convexity may fail. These are exactly the failure modes one expects from EM+BIC. Theorem 2 stays exact regardless. The audits in §6 and App. J.4 measure the leakage directly.

Remark 6 (Copula component). The Gaussian-copula component admits an analogous argument after quantile clipping. Replacing one observation changes each empirical CDF by at most $1/N_c$. Clipping normal scores to $[\varepsilon, 1-\varepsilon]$ makes $u \mapsto \Phi^{-1}(u)$ Lipschitz with constant $1/\phi(\Phi^{-1}(\varepsilon))$, so the normal-score matrix changes by $O(1/N_c)$. The Ledoit–Wolf covariance is a smooth shrinkage functional of the empirical covariance. Under (ii), the induced Gaussian KL is quadratic in the covariance perturbation. The copula component is therefore replace-one stable at order $O(1/N_c)$ in parameters and $O(1/N_c^2)$ in one-sample KL, conditional on clipping and covariance regularization. This statement covers the latent Gaussian dependence estimator only. The released features pass through the data-dependent inverse empirical marginals of Algorithm 3, whose support can change when an extreme observation is replaced, so replace-one KL stability of the released feature distribution would need an additional common-support or smoothing assumption that we do not establish.

Empirical stability diagnostic. We measure the GMM-component proxy $\widehat{\Gamma}_{\mathrm{gen}}^{\mathrm{GMM}} := m \cdot \max_i \max\big(\widehat{\mathrm{KL}}(G_{\mathcal{D}} \| G_{\mathcal{D}^{(i)}}), \widehat{\mathrm{KL}}(G_{\mathcal{D}^{(i)}} \| G_{\mathcal{D}})\big)$, where $\mathcal{D}^{(i)}$ replaces z_i with a uniformly chosen same-class record (30 perturbations per dataset, Monte Carlo $\widehat{\mathrm{KL}}$ with $S = 2000$ samples in each direction, $\alpha = 2$, $m = \alpha N$). Table 5 reports the result on the full 14-dataset numeric tier.

First, the BIC-selected component count is stable on $\ge 77\%$ of replace-one perturbations on every dataset (mean 0.95, the lowest values, on Pendigits, Banknote, and Diabetes, are large-N or low-d

regimes where one swap can flip a borderline BIC choice), supporting the BIC-margin part of the local-stability conditions of Proposition 7. Second, the per-record KL is small (≤ 0.05 nats) on $8/14$ datasets, moderate (0.6–1.4) on three, and larger (2.6–5.2) on three high-d or many-class settings (Segment, Pendigits, Dermatology) where a single record can shift one component's covariance materially. Third, the m-factor pushes the cumulative $\widehat{\Gamma}_{\text{gen}}^{\text{GMM}}$ into the regime where Pinsker's $\sqrt{\widehat{\Gamma}/8}$ ceases to be a tight numerical bound on advantage on most datasets. The audited NN-MIA AUC of Table 9 stays at 0.49–0.60 throughout, consistent with the looseness of Pinsker for large total KL rather than with a stability failure. For the Copula component, Remark 6 gives a $O(1/N_c)$-scale parameter perturbation argument. When self-prediction selection is enabled, direct density evaluation of $Q_{\mathcal{D}}$ is unavailable, so we complement this generator-level diagnostic with the $K{=}1$ vs. $K{=}10$ audit of Table 40. These numbers are a sampled maximum over 30 perturbations of the GMM component, not the supremum over neighbours of the full release law that Γ_∞ requires, so they are a diagnostic rather than a bound.

Table 5: Empirical $\widehat{\Gamma}_{\text{gen}}^{\text{GMM}}$ on the full 14-dataset numeric tier (30 replace-one perturbations, $S{=}2000$ MC samples, $\alpha{=}2$). Per-record KL is the maximum over both directions, and "same-k" is the fraction of perturbations that keep the BIC-selected component count for every class, providing an empirical component-count stability diagnostic for assumption (iii) of Proposition 7. Same-k holds with mean 0.95 (min 0.77) across all 14 datasets. The Pinsker-implied advantage bound $\sqrt{\widehat{\Gamma}/8}$ is vacuous (>0.5) on most datasets because of the m-factor. The empirically audited NN-MIA AUC of Table 9 stays at 0.49–0.60 throughout, consistent with Pinsker looseness rather than a stability failure.

Dataset	N	d	m	**same-k**	**KL/rec (max)**	$\widehat{\Gamma}_{\text{gen}}^{\text{GMM}}$
Iris	150	4	300	1.00	0.0096	2.87
Wine	178	13	356	1.00	0.0340	12.11
BreastCancer	569	30	1138	1.00	0.0039	4.48
Diabetes	768	8	1536	0.83	1.3723	2107.84
Vehicle	846	18	1692	1.00	0.0132	22.28
Segment	2310	18	4620	0.97	3.6846	17022.70
Banknote	1372	4	2744	0.80	0.6305	1730.21
Climate-Model	540	20	1080	1.00	0.0131	14.17
Ionosphere	351	34	702	1.00	0.9334	655.26
Seeds	210	7	420	1.00	0.0044	1.86
Dermatology	358	34	716	1.00	5.2233	3739.86
Pendigits	10992	16	21984	0.77	2.6451	58150.76
Waveform	5000	40	10000	1.00	0.0008	8.35
Optdigits	1797	64	3594	1.00	0.0243	87.36
14-tier mean				0.95	1.04	5969

D Algorithm Details

D.1 BIC-Optimal GMM: Full Pseudocode

Algorithm 2 fits a per-class Gaussian mixture with the component count selected by the Bayesian Information Criterion. Lines 1–4 partition the training set by class and choose the candidate-count cap $k_c^{\max}$ with the degeneracy guard from Leroux (1992). Lines 5–12 sweep $k = 1, \ldots, k_c^{\max}$, fitting a k-component full-covariance GMM via Expectation–Maximization with regularization ρ, scoring each fit by $\text{BIC}(k) := -2 \log \hat{L}_k + n_p(k) \log N_c$ where $n_p(k)$ is the number of free parameters of a k-component full-covariance GMM in d dimensions, and retaining the lowest-BIC fit.

Algorithm 2 BIC-Optimal GMM: Fit

Require: Training data $X \in \mathbb{R}^{N \times d}$, labels $y \in [C]^N$, $k_{\max} \in \mathbb{Z}_{>0}$, regularization $\rho > 0$
Ensure: Per-class models $\{\hat{g}_c\}_{c=1}^{C}$, priors $\{\hat{\pi}_c\}$

for $c = 1, \ldots, C$ **do**
  $X_c \leftarrow \{x_i : y_i = c\}$; $N_c \leftarrow |X_c|$
  $k_c^{\max} \leftarrow \max(1, \min(k_{\max}, \lfloor N_c/(d+2) \rfloor))$
  best_bic $\leftarrow +\infty$
  for $k = 1, \ldots, k_c^{\max}$ **do**
    Fit GMM with k full-covariance components, reg ρ, via EM
    $\mathrm{BIC}_k \leftarrow -2 \log \hat{L}_k + n_p(k) \log N_c$
    if $\mathrm{BIC}_k <$ best_bic **then**
      best_bic $\leftarrow \mathrm{BIC}_k$; $\hat{g}_c \leftarrow \mathrm{GMM}(k)$
    end if
  end for
  $\hat{\pi}_c \leftarrow N_c/N$
end for

D.2 Gaussian Copula: Full Pseudocode

Algorithm 3 fits and samples a per-class Gaussian copula. The fit phase (lines 1–9) computes a quantile grid $q_{c,j}$ at Q evenly spaced probabilities for each feature j, stores the sorted feature values $\mathrm{sv}_{c,j}$ for inverse-CDF lookup, and converts ranks to standard-normal scores $Z_{i,j}$.

Algorithm 3 Gaussian Copula Synthesizer: Fit and Sample

Require: $X_c \in \mathbb{R}^{N_c \times d}$, quantile resolution $Q = 200$
Ensure: Copula parameters $(\hat{\Sigma}_c, \hat{F}_{c,1}, \ldots, \hat{F}_{c,d})$

for $j = 1, \ldots, d$ **do**
  $q_{c,j} \leftarrow$ quantile grid of $X_{c,j}$ at Q evenly spaced probabilities
  Store sorted values $\mathrm{sv}_{c,j}$
  $u_{i,j} \leftarrow (\mathrm{rank}(x_{i,j}) - 0.5)/N_c$; $Z_{i,j} \leftarrow \Phi^{-1}(u_{i,j})$
end for
if $N_c < 3d$ **then**
  $\hat{\Sigma}_c \leftarrow$ LedoitWolf . fit(Z_c)
else
  $\hat{\Sigma}_c \leftarrow \mathrm{corr}(Z_c)$
end if

Sample(n_s)**:**
Draw $Z^* \sim \mathcal{N}(0, \hat{\Sigma}_c)^{n_s}$
for $j = 1, \ldots, d$ **do**
  $u_j^* \leftarrow \Phi(Z_j^*)$
  $X_j^* \leftarrow$ interpolate$(\mathrm{sv}_{c,j}, q_{c,j}, u_j^*)$
end for
return X^*

E Context-Sensitivity Diagnostic

We probe the context-Lipschitz behavior of f_θ (Assumption 1) by replacing one real context record with a synthetic draw and measuring the ratio $\|\Delta\mathrm{pred}\|_\infty / \bar{W}_1$, where $\|\Delta\mathrm{pred}\|_\infty$ is the ℓ_∞ change in TabPFN's predicted class-probability vector and $\bar{W}_1$ is the average 1-Wasserstein distance between the original and perturbed empirical contexts. The ratio is a Monte-Carlo lower-bound estimator of L_f from Assumption 1. Across 264 trials on all six core datasets, the median ratio is 20.3 and the 95th percentile p_{95} is 112.8 (Table 6). Ratios are finite in every trial, consistent with bounded sensitivity but not a formal proof.

Table 6: Context-sensitivity diagnostic. 3-fold, 15 perturbations per fold.

Dataset	**Median**	**Mean**	p_{95}	n
Iris	71.6	122.8	446.6	45
Wine	0.8	1.0	2.5	45
BreastCancer	1.2	1.8	5.0	45
Multi3_Med	27.9	32.3	71.1	45
Multi5_Med	24.3	26.3	48.4	39
HighDim_50	25.8	25.7	46.1	45
Overall	20.3	35.2	112.8	264

F Additional Experimental Results

Table 7: SDS vs. ForestDiffusion on the 14-dataset numeric tier (5-fold $\times$ 3 seeds, α=2) and the core-six tier (5-fold $\times$ 5 seeds). Arrows indicate metric direction. Advantage (parentheses) is the chance-centered AUC gap of Eq. (2), and QAPD-MIA is the query-adaptive attack under Membership leakage. Table 10 uses the separate K-RFF geometry-attack protocol, under which SDS's core-six mean is 0.027.

Metric	**SDS**	**ForestDiffusion**
Per-dataset accuracy wins, 14-tier ↑	6/14	8/14
Per-dataset NN-MIA wins, 14-tier ↑	**14/14**	0/14
Mean NN-MIA AUC (advantage), 14-tier ↓	**0.54** (**0.04**)	0.73 (0.23)
Mean QAPD-MIA AUC (advantage), 14-tier ↓	**0.56** (**0.06**)	0.72 (0.22)
Per-dataset QAPD wins, 14-tier ↑	**14/14**	0/14
Attribute R^2, real-data core ↓	0.77	**0.76**

F.1 Hardening Noise Mechanism

The hardened generator adds noise at three points, each perturbing one population statistic the attribute attacker relies on: Laplace noise on the class counts (ε=1, counts have sensitivity 2), Gaussian noise with standard deviation 0.3 on the copula covariance, and Gaussian noise on the GMM means calibrated as in differential privacy at $(\varepsilon, \delta)=(1, 10^{-5})$.

F.2 Practical Significance and Mitigation of Attribute Inference

A large share of population-level attribute structure is recoverable from any useful synthetic release, ours included, and this is not fully solvable within the pipeline. It becomes a practical risk specifically when an adversary already holds a target's quasi-identifiers (attributes such as age or zip code that narrow a person down without naming them) and the sensitive column correlates strongly with those quasi-identifiers at the population level. BreastCancer's radius and perimeter columns are the extreme case ($R^2 > 0.99$). A column with $R^2 \approx 0.5$ that correlates only weakly with the available quasi-identifiers poses much less risk in practice. Membership, geometry, and near-record leakage do not follow this pattern, so attribute inference does not by itself identify which record was real. Mitigations beyond the hardening of §6: suppressing or noising specific regulated columns before synthesis is compatible with the pipeline and needs no change to the generator, and simply not releasing columns the downstream task does not need is always available.

Deployment recipe for regulated columns. We recommend a per-column policy. (i) Audit each sensitive column's R^2_{attr} against the quasi-identifiers actually available to the adversary, not against every other column. (ii) For any column above a chosen threshold (e.g. $R^2_{\text{attr}} > 0.5$), either suppress it, coarsen it, or increase the hardening noise (the label-count budget ε_ℓ, the covariance noise σ, and the mean budget ε_{dp} of App. F.1) until the audited R^2_{attr} falls below the threshold, accepting the resulting utility cost. (iii) Never release columns the downstream task does not need. We propose, but have not evaluated, setting these by sweeping them on an owner-side validation split and taking the smallest noise that meets the target. The hardening results of §6 use fixed values (ε_ℓ=1, σ=0.3, ε_{dp}=1), with

the two-point σ ablation of Table 21 the only sweep performed. Because per-column R^2 measures marginal recoverability rather than joint reconstructability, it should be read alongside rarity-weighted risk (leakage concentrates on rare records) and a memorization-versus-population diagnostic that compares a train-target attacker to a population-matched one, following the reconstruction-attack systematization of Golob et al. (2026): a large marginal R^2 that is fully explained by population conditionals reflects learnable structure, not memorization of an individual.

Compute. Experiments ran on a Slurm cluster whose nodes each provide $8\times$ NVIDIA B200 GPUs (180 GB HBM each), dual Intel Xeon Platinum 8570 CPUs (160 threads), and 1.8 TB RAM. Every individual evaluation fits on a single GPU. The 2026-generation study (App. L.7) shards datasets across GPUs in parallel. Earlier rounds used a single NVIDIA A100 (80 GB) node. The headline TabPFN reader is our standalone re-implementation of the TabPFN-2.5 architecture loading the official weights (24 layers, 10.7M parameters), shipped in the supplementary code, and the TabPFN-2.5 row of Table 3 runs the official package on the same weights.

F.3 Per-Dataset Membership and Geometry Audit

Table 8: Synthesizer ablation on the core-six tier (5-fold $\times$ 5 seeds). Δ_{util} is the synthetic-context minus real-context accuracy gap. Closer to zero is better. w_{GMM} is the ensemble weight on the GMM component.

Synthesizer	Δ_{util} (%)
GMM-k=2 (fixed, no BIC)	-2.0
BIC-GMM only (w_{GMM}=1)	-0.5
Copula only (w_{GMM}=0)	-0.4
BIC-GMM + Copula 50:50 (default)	$\mathbf{-0.2}$

Table 9: Per-dataset NN-MIA AUC on the 14-dataset headline tier (5-fold $\times$ 3 seeds, α=2). Chance is 0.5. FDiff is ForestDiffusion.

Dataset	**NN-MIA AUC** $\downarrow$	
	SDS	**FDiff**
Iris	**0.60**	0.88
Wine	**0.55**	0.78
BreastCancer	**0.53**	0.66
Diabetes	**0.53**	0.74
Vehicle	**0.53**	0.74
Segment	**0.52**	0.65
Banknote	**0.51**	0.77
Ionosphere	**0.53**	0.58
Seeds	**0.49**	0.85
Dermatology	**0.57**	0.71
Pendigits	**0.55**	0.72
Waveform	**0.55**	0.74
Optdigits	**0.57**	0.68
Climate-Model	**0.55**	0.77
Mean	**0.54**	0.73

Table 10: Geometry-attack R^2 (MDS + Procrustes alignment, §2, per-tier encoding dimension and σ in App. F.4). The columns are not comparable: K-RFF encodes the real records row for row so the attacker holds the true correspondence, while a synthetic release admits none and is scored under an arbitrary index pairing. Under that pairing a row-permuted copy of the real records scores what SDS scores (App. L.10).

Dataset	R^2_{geom} $\downarrow$	
	K-RFF	**SDS**
Iris	0.89	0.01
Wine	0.95	0.04
BreastCancer	0.23	0.01
Multi3_Med	0.49	0.02
Multi5_Med	0.92	0.03
HighDim_50	0.94	0.06
Mean	0.735	0.027

F.4 K-RFF Baseline Specification

K-RFF is our concrete instance of a *distance-preserving encoding*: the class of obfuscations designed to preserve accuracy (random Fourier features (Rahimi & Recht, 2007) and Johnson–Lindenstrauss random projections (Blocki et al., 2012)) whose defining property is that pairwise distances survive the transform, so a downstream in-context model still sees a faithful geometry. The keyed random Fourier feature (K-RFF) baseline encodes each record $x \in \mathbb{R}^d$ as $z(x) = \sqrt{2/D}\left[\cos(\omega_i^\top x + b_i)\right]_{i=1}^{D}$

with $\omega_i \sim \mathcal{N}(0, \sigma^{-2} I_d)$ and $b_i \sim \mathrm{Unif}[0, 2\pi]$ (Rahimi & Recht, 2007), where the projection (Ω, b) is generated pseudorandomly from a secret key shared between the data owner and the prediction service. The encoding hides raw feature values while approximately preserving pairwise distances (it approximates a Gaussian kernel), which is what is meant to preserve in-context accuracy and, simultaneously, what makes the layout of the encoded records recoverable by the MDS-plus-Procrustes attack of §2. The per-tier encoding dimension D and bandwidth σ used in Table 10, together with the full encoder and attack implementations, are provided in the supplementary code (`attack_geometry.py`).

Unequal-size alignment protocol. Against K-RFF, the released matrix encodes the N real records row-for-row, so the attacker holds the true one-to-one correspondence: the attack inverts the kernel to squared distances, embeds by metric MDS, and Procrustes-aligns (rotation, reflection, and scale) to the standardized real records. Against a synthetic release of size $m \neq N$, no record correspondence exists. The audit therefore subsamples both sides to a common size $n = \min(N, m, 100)$, embeds the synthetic subsample by metric MDS on its pairwise distances, and Procrustes-aligns to the real subsample under an index pairing. R^2_{geom} is computed on the aligned pair. This evaluates the strongest rotation/reflection/scale alignment given an arbitrary pairing. Attacks that jointly optimize the correspondence (e.g. Gromov–Wasserstein matching) are outside the evaluated family and are listed as an open extension in App. M.

F.5 Per-Dataset Headline-Tier Accuracy

Table 11: Ensemble on 14 public benchmarks. 5-fold × 3 seeds. Δ = Acc-Synth − Acc-Real (%). N is the row count after this protocol's size cap.

Dataset	N	d	C	Δ	Std
Iris	150	4	3	−0.4	2.1
Wine	178	13	3	−0.4	1.7
BreastCancer	569	30	2	−1.2	1.8
Diabetes	768	8	2	+0.4	2.4
Vehicle	846	18	4	−5.7	2.4
Segment	2 000	19	7	−4.4	1.8
Banknote	1 372	4	2	−0.3	0.4
Climate-Model	540	20	2	+0.1	1.2
Ionosphere	351	34	2	−1.4	2.7
Seeds	210	7	3	−0.2	2.2
Dermatology	358	34	6	−0.0	1.4
Pendigits	1 000	16	10	−1.2	1.2
Waveform	1 000	40	3	−2.3	1.9
Optdigits	1 000	64	10	−2.1	1.4
Mean				**−1.4**	

Across all 14 headline-tier datasets, compared on identical folds and identical size-capped data (App. F.20), SDS+TabPFN matches or exceeds Random Forest on $10/14$ ($+1.0\%$ mean). Three of the four Random Forest wins (Diabetes, Climate-Model, Optdigits) are within noise. The one substantive loss is Segment (-2.9pp), which carries the second-largest synthesis gap on this tier after Vehicle.

F.6 Robust Statistics on the Headline Tier

Table 12 reports distribution-free statistics for the 14-dataset repeated tier: bootstrap and paired-t confidence intervals on the mean and median gap, one-sided non-inferiority tests (the lower half of the two-one-sided procedure of Schuirmann, 1987) at three margins, and per-dataset coverage and tail statistics. The test at δ=2.5% bounds the mean 14-tier gap below 2.5% (p=0.016).

Table 12: Robust statistics on the 14-dataset repeated tier (Table 11). The bootstrap CIs make no normality assumption. The multi-margin test is a one-sided non-inferiority test (null $\bar{\Delta} \leq -\delta$), not two-one-sided equivalence, and it rejects at δ=2.5% but not at δ=1%, making the dependence on the chosen margin explicit. Per-dataset coverage and tail stats bound how badly any single dataset can degrade. All values in percentage points.

Statistic	**Value**
Mean Δ	-1.4
Median Δ	-0.8
Bootstrap 95% CI on mean (B=20,000)	$[-2.3, -0.6]$
Bootstrap 95% CI on median	$[-1.9, -0.2]$
Paired-t 95% CI on mean	$[-2.4, -0.3]$
Worst case (Vehicle)	-5.7
90th-percentile worst case	-3.8
One-sided non-inferiority p at δ=1%	0.77
One-sided non-inferiority p at δ=2.5%	0.016^{*}
One-sided non-inferiority p at δ=5%	$<10^{-3*}$
Datasets within $\lvert\Delta\rvert \leq 1\%$	7/14
Datasets within $\lvert\Delta\rvert \leq 2.5\%$	12/14
Datasets within $\lvert\Delta\rvert \leq 5\%$	13/14

*rejects H_0: $\bar{\Delta} \leq -\delta$ at α=0.05.

Table 13: Mixed-type benchmark: accuracy gap (Δ) between Mixed-Type SDS synthetic context and real training context on 15 CC18/OpenML datasets, 3-fold protocol. Entries marked † are the all-categorical datasets, evaluated under the discrete pipeline of §4. A contingency-table synthesizer that enumerates training rows reaches lower $\lvert\Delta\rvert$ on them at MIA AUC ≈ 0.93, whereas the discrete pipeline keeps MIA at chance. *Cat* is the number of categorical columns.

Dataset	N	d	Cat	Δ
Tic-Tac-Toe†	958	9	9	-8.2
Car-Eval†	1728	6	6	-1.4
Kr-vs-Kp†	2000	36	36	0.0
Nursery†	2000	8	8	-0.5
Balance-Scale†	625	4	4	-8.3
Adult	2000	14	8	-2.5
Splice†	2000	60	60	-1.1
Electricity	2000	8	1	-3.5
Mushroom†	2000	22	22	-1.2
CMC	1473	9	7	-2.4
Hypothyroid	2000	29	23	-1.2
Yeast	1484	8	2	-2.3
German-Credit	1000	20	17	-2.5
Bank-Marketing	2000	16	9	-0.3
Anneal	898	38	35	-3.8
Mean				-2.6

F.7 Per-Dataset Mixed-Type Accuracy

The two largest gaps in Table 13, Tic-Tac-Toe (-8.2%) and Balance-Scale (-8.3%), arise because the labels in these datasets are strict higher-order combinatorial functions of the features (three-in-a-row patterns, weight-times-distance imbalance), and a pairwise Chow-Liu tree augmented with a label refiner cannot fully recover such rules from ≤ 500 samples per class. A contingency-table synthesizer that enumerates training rows verbatim closes these gaps, but at MIA AUC ≈ 0.93. The discrete pipeline of §4 keeps MIA at chance ($[0.49, 0.57]$, Table 17) at the cost of the larger accuracy gap.

Table 14: Obfuscation audit on the core six. AUC orientation: higher = stronger adversary. Distributional discrimination (logistic regression, neural-net), nearest-neighbor MIA on a held-out fold (MIA-Hld), shadow-model MIA, and marginal Wasserstein-1 distance.

Dataset	**Log. Disc.**	**Neur. Disc.**	**MIA-Hld**	**Shadow**	W_1
Iris	0.549	0.512	0.524	0.541	0.051
Wine	0.562	0.521	0.507	0.545	0.058
BreastCancer	0.556	0.509	0.520	0.514	0.072
Multi3_Med	0.551	0.518	0.519	0.532	0.061
Multi5_Med	0.564	0.519	0.526	0.549	0.069
HighDim_50	0.565	0.511	0.518	0.570	0.095
Mean	0.558	0.515	0.519	0.542	0.068

F.8 Core-Six Obfuscation Audit

F.9 Modern Baselines (TVAE)

Table 15: SDS Ensemble vs. TVAE on 14 public benchmarks (3-fold $\times$ 2 seeds). Δ = Synth $-$ Real (%). SDS wins 13/13 where both succeed. Wilcoxon $p = 0.00012$, Cohen's $d = 1.18$. Both means are over the 13 datasets where TVAE succeeds, so they exclude Climate-Model.

Dataset	N	**Δ-SDS**	**Δ-TVAE**
Iris	150	$\mathbf{+0.0}$	-2.3
Wine	178	$\mathbf{-0.8}$	-2.5
BreastCancer	569	$\mathbf{-1.1}$	-4.4
Diabetes	768	$\mathbf{-1.1}$	-6.0
Vehicle	846	$\mathbf{-4.8}$	-37.9
Segment	2 000	$\mathbf{-4.0}$	-10.3
Banknote	1 372	$\mathbf{-0.4}$	-13.9
Ionosphere	351	$\mathbf{+0.0}$	-11.7
Seeds	210	$\mathbf{+0.2}$	-7.9
Dermatology	358	$\mathbf{-0.3}$	-25.8
Pendigits	1 000	$\mathbf{-1.5}$	-26.3
Waveform	1 000	$\mathbf{-2.4}$	-7.1
Optdigits	1 000	$\mathbf{-2.0}$	-27.3
Climate-Model	540	$\mathbf{+0.1}$	fail
Mean		$\mathbf{-1.4}$	-14.1

F.10 Baseline Tuning Sweep

To verify that the deep-baseline gap is not driven by default hyperparameters alone, we swept configurations for TVAE (epochs $\in \{200, 600\}$, embedding dim $\in \{64, 128, 256\}$) and for CTGAN (epochs $\in \{300, 600\}$, batch size $\in \{100, 500\}$, generator depth $\in \{128, 256\}$) on a four-dataset subset of the repeated tier under 3-fold $\times$ 2 seeds. The sweep suggests the gap is not purely due to default hyperparameters, and we do not generalize it beyond the four datasets evaluated. The one case where tuned TVAE matches SDS is BreastCancer (TVAE-600ep+latent64 $\Delta = -0.9\%$ vs. SDS -1.2%), a binary task. Multi-class and weak-class-separation datasets in the subset show larger gaps.

F.11 Broader-Tier ForestDiffusion MIA Audit

The 14-dataset headline-tier comparison is reported in Table 9. The broader audit on the 15 CC18/OpenML mixed-type benchmarks and 6 additional numeric benchmarks is reported below.

F.12 Label-Aware MIA on Categorical Benchmarks

The audit so far reports the standard Stadler nearest-neighbor MIA, which only uses synthetic features X_{synth}. An adversary that also receives synthetic labels y_{synth} could in principle exploit label co-

Table 16: Baseline tuning sweep: best-per-dataset tuned Δ vs. SDV defaults. SDS column for reference. 3-fold $\times$ 2 seeds.

Dataset	**Real**	**SDS**	TVAE-def.	TVAE-best	CTGAN-def.	CTGAN-best
Iris	96.7	$\mathbf{+0.7}$	-3.0	0.0	-62.0	-6.7
BreastCancer	97.5	-1.2	-4.2	$\mathbf{-0.9}$	-29.4	-6.3
Vehicle	85.1	$\mathbf{-4.1}$	-36.3	-18.6	-61.6	-41.8
Pendigits	98.5	$\mathbf{-1.4}$	-24.5	-8.4	-91.9	-26.1
Mean		$\mathbf{-1.5}$	-17.0	-7.0	-61.2	-20.2

Table 17: Broad-tier MIA audit: SDS vs. ForestDiffusion on 21 datasets covering the 15 CC18/OpenML mixed-type benchmarks and 6 additional numeric benchmarks, 2-fold protocol. AUC reported; adversarial advantage is $\max(\text{AUC}, 1-\text{AUC}) - 0.5$. SDS advantage ≤ 0.07 per dataset; FDiff exceeds advantage 0.10 on 10 of 21. Combined with the 14 repeated-tier datasets in Table 9, SDS wins on 35/35. Entries marked † are the all-categorical datasets evaluated under the discrete pipeline of §4.

Dataset	N	d	**SDS**	**FDiff**
15 CC18/OpenML mixed-type benchmarks				
Tic-Tac-Toe†	958	9	$\mathbf{0.54}$	0.70
Car-Eval†	1 728	6	$\mathbf{0.51}$	0.64
Kr-vs-Kp†	2 000	36	$\mathbf{0.51}$	0.56
Nursery†	2 000	8	$\mathbf{0.52}$	0.67
Balance-Scale†	625	4	$\mathbf{0.54}$	0.77
Mushroom†	2 000	22	$\mathbf{0.51}$	0.62
Adult	2 000	14	$\mathbf{0.50}$	0.58
Splice†	2 000	60	$\mathbf{0.51}$	0.61
Electricity	2 000	8	$\mathbf{0.52}$	0.66
CMC	1 473	9	$\mathbf{0.51}$	0.55
Hypothyroid	2 000	29	$\mathbf{0.49}$	0.57
Yeast	1 484	8	$\mathbf{0.52}$	0.75
German-Credit	1 000	20	$\mathbf{0.51}$	0.57
Bank-Marketing	2 000	16	$\mathbf{0.50}$	0.58
Anneal	898	38	$\mathbf{0.51}$	0.57
6 additional numeric benchmarks				
Spambase	2 000	57	$\mathbf{0.51}$	0.60
Pc4	1 458	37	$\mathbf{0.50}$	0.56
Kc1	2 000	21	$\mathbf{0.50}$	0.53
Blood-Transfusion	748	4	$\mathbf{0.52}$	0.65
Phoneme	2 000	5	$\mathbf{0.51}$	0.72
Sonar	208	60	$\mathbf{0.57}$	0.60
Mean			$\mathbf{0.51}$	0.62

occurrence, particularly on datasets where the label is a deterministic rule (Tic-Tac-Toe), an arithmetic of the features (Balance-Scale), or a high-information categorical (Mushroom, Nursery, Kr-vs-Kp). We extend the audit with three label-aware attacks against $(X_{\text{synth}}, y_{\text{synth}})$:

- **NN-L:** class-conditional nearest neighbor. The target (x^*, y^*) scores $-\min_{j:\, y^s_j = y^*} \|x^* - x^s_j\|_2$, falling back to a global nearest neighbor when no synthetic record carries label y^*.
- **CLF-L:** a four-feature logistic regression with 3-fold CV: global DCR, class-conditional DCR, label agreement at the nearest neighbor, and fraction of the K_{NN}=5 nearest neighbors whose label matches.
- **QAPD-L:** the query-adaptive prediction-perturbation attack of Stadler et al. (2022) augmented with the TabPFN posterior at the *true* label and its stability under input perturbation.

We run 3-fold CV on the five most label-informative categorical datasets. Table 18 reports AUC, with advantage in parentheses. Even the strongest label-aware attack against SDS stays ≤ 0.06 on every

dataset (mean 0.018): labels give the adversary no meaningful additional signal. ForestDiffusion's leakage is not suppressed under label-aware attacks (advantage 0.12–0.15 across the four attacks).

Table 18: Label-aware MIA on 5 categorical datasets. AUC reported with adversarial advantage $\max(\text{AUC}, 1-\text{AUC})-0.5$ in parentheses. NN-feat is the feature-only Stadler attack; NN-L, CLF-L, and QAPD-L additionally use synthetic labels. SDS stays within 0.06 of chance under every attack on every dataset; FDiff exceeds 0.18 on three of five.

Dataset	Method	NN-feat	NN-L	CLF-L	QAPD-L
Tic-Tac-Toe	SDS	**0.53** (0.03)	**0.52** (0.02)	**0.49** (0.01)	**0.52** (0.02)
	FDiff	0.69 (0.19)	0.68 (0.18)	0.66 (0.16)	0.71 (0.21)
Balance-Scale	SDS	**0.53** (0.03)	**0.56** (0.06)	**0.54** (0.04)	**0.52** (0.02)
	FDiff	0.72 (0.22)	0.75 (0.25)	0.72 (0.22)	0.69 (0.19)
Mushroom	SDS	**0.51** (0.01)	**0.50** (0.00)	**0.50** (0.00)	**0.53** (0.03)
	FDiff	0.59 (0.09)	0.59 (0.09)	0.59 (0.09)	0.48 (0.02)
Kr-vs-Kp	SDS	**0.50** (0.00)	**0.50** (0.00)	**0.48** (0.02)	**0.51** (0.01)
	FDiff	0.55 (0.05)	0.56 (0.06)	0.55 (0.05)	0.50 (0.00)
Nursery	SDS	**0.51** (0.01)	**0.50** (0.00)	**0.49** (0.01)	**0.49** (0.01)
	FDiff	0.64 (0.14)	0.65 (0.15)	0.65 (0.15)	0.69 (0.19)
Mean, SDS		0.52 (0.015)	0.52 (0.018)	0.50 (0.017)	0.51 (0.018)
Mean, FDiff		0.64 (0.139)	0.65 (0.145)	0.63 (0.134)	0.61 (0.122)

F.13 Broad Attribute-Inference Audit

Because attribute inference is the dominant residual leakage and the core-six audit covers only six datasets, we extend it to 35 datasets across three tiers and stratify by feature type:

- *Tier A (14 repeated-tier, numeric):* Iris, Wine, BreastCancer, Diabetes, Vehicle, Segment, Banknote, Climate-Model, Ionosphere, Seeds, Dermatology, Pendigits, Waveform, Optdigits.
- *Tier B (15 mixed-type):* Tic-Tac-Toe, Car-Eval, Kr-vs-Kp, Nursery, Balance-Scale, Adult, Splice, Electricity, Mushroom, CMC, Hypothyroid, Yeast, German-Credit, Bank-Marketing, Anneal. Hypothyroid failed sampling for SDS in the discrete pipeline (rare class) and is reported as `n/a`; we keep $n{=}14$ for the tier mean.
- *Tier C (TabZilla representative, numeric):* Spambase, Pc4, Kc1, Blood-Transfusion, Phoneme, Sonar.

For each dataset we fit the synthesizer (Mixed-Type for tier B, Ensemble for tiers A and C), draw $2\times$ synthetic, and per column j either fit Ridge from $X_{\text{synth},-j}$ to $X_{\text{synth},j}$ for numeric j or a HistGradientBoosting classifier for categorical j, then evaluate on the real data. We report (i) *num→num* mean Ridge R^2, (ii) *cat→cat* mean accuracy gain over the majority baseline, (iii) *overall* mean across all columns, and (iv) the *sensitive worst column*, defined as the largest single-column leakage and corresponding to the realistic threat where a particular column is sensitive. For numeric columns where SDS sampling produced a few extreme outliers (Segment, Ionosphere, Optdigits) we clip synthetic values to the real $[1, 99]$ percentile range before fitting Ridge. This represents an outlier-resilient attacker that discards implausible points and yields a tighter estimate of leakage. ForestDiffusion is run wherever feasible. We skip when $d{>}35$ or training exceeds 600 s.

Table 19 reports tier means and Table 20 the per-dataset breakdown for tier B (mixed-type). The results reinforce the core paper claim. (1) On numeric datasets the SDS and FDiff R^2 are essentially tied. This is the utility–leakage coupling: any generator that preserves the correlations TabPFN exploits will also support attribute inference at roughly that fidelity. (2) On mixed-type datasets the categorical attacker gains slightly more from SDS than from FDiff (cat-ACC* 0.04 vs. -0.03), consistent with SDS preserving the categorical structure that FDiff often samples near-randomly. This is the utility–leakage coupling again rather than an advantage, and because the two methods cover different subsets the tier means are not matched. The DP-compatible hardening brings the real-data core to 0.52 (Table 21) and the tier-A mean to 0.48 (Table 22) if a deployment requires a stricter bound.

Table 19: Broad attribute-inference summary across three tiers (34 datasets; Hypothyroid excluded for SDS due to a rare-class sampling failure in the discrete pipeline). Numbers are mean R^2 (numeric) or mean accuracy gain over majority baseline (categorical). *Sens.* is the mean over per-dataset *worst* columns. FDiff covers fewer datasets where $d > 35$ or training exceeded 600 s.

Tier	**Method**	**num→num** R^2	**cat→cat** ACC*	**overall**		**sens. max**
A (14 numeric repeated)	SDS	0.59	–	0.59		0.89
	FDiff	0.62	–	0.62	0.89	(n=11)
B (15 mixed-type)	SDS	0.25	**0.04**	0.08		0.29
	FDiff	0.27	−0.03	0.04	0.29	(n=10)
C (6 TabZilla numeric)	SDS	0.57	–	0.57		0.85
	FDiff	0.55	–	0.55	0.73	(n=3)

Table 20: Per-dataset attribute-inference on the 15 mixed-type tier (Hypothyroid excluded for SDS). *num* is mean Ridge R^2 over numeric columns (– when none); *cat* is mean accuracy gain over majority over categorical columns; *sens* is the worst single column. On categorical columns the attacker gains more from SDS than from FDiff (cat-mean $+0.04$ against -0.03), i.e. SDS leaks more on this axis, not less. The two methods also cover different dataset subsets, so the tier means are not matched, and categorical accuracy gain and numeric R^2 are reported separately rather than combined.

				SDS			**FDiff**		
Dataset	N	d	d_{cat}	num	cat	sens	num	cat	sens
Tic-Tac-Toe	958	9	9	–	−0.01	0.03	–	−0.04	−0.02
Car-Eval	1 728	6	6	–	0.00	0.00	–	0.00	0.00
Kr-vs-Kp	2 000	36	36	–	0.04	0.36	–	–	–
Nursery	2 000	8	8	–	0.01	0.02	–	0.02	0.05
Balance-Scale	625	4	4	–	0.00	0.00	–	0.00	0.00
Adult	2 000	14	8	0.07	0.03	0.16	0.08	−0.12	0.19
Splice	2 000	60	60	–	0.03	0.08	–	–	–
Electricity	2 000	8	1	0.37	0.18	0.66	0.37	0.14	0.66
Mushroom	2 000	22	22	–	0.19	0.49	–	–	–
CMC	1 473	9	7	0.36	0.03	0.38	0.37	−0.11	0.39
Hypothyroid	2 000	29	23	n/a (rare-class fail)			–	–	–
Yeast	1 484	8	2	0.16	0.00	0.38	0.16	−0.01	0.37
German-Credit	1 000	20	17	0.42	0.01	0.53	0.42	−0.11	0.54
Bank-Marketing	2 000	16	9	0.17	0.02	0.52	0.21	−0.03	0.74
Anneal	898	38	35	0.19	0.05	0.40	–	–	–
Mean				0.25	**0.04**	0.29	0.27	−0.03	0.29

F.14 Obfuscation Hardening Ablation

Table 21: Obfuscation hardening: mean over Iris, Wine, BreastCancer (3-fold × 3 seeds). Δ^{+}_{SDS}: additional accuracy cost beyond the unhardened SDS context (not the real context).

Configuration	Δ^{+}_{SDS}	**MIA**	**Attr** R^2	W_1
Baseline (Ensemble)	NA	0.553	0.770	0.082
A: Label-priv. (ε_ℓ=1)	−0.1	0.544	0.769	0.091
B: CorrNoise (σ=0.1)	−0.4	0.559	0.742	0.087
B: CorrNoise (σ=0.3)	−0.8	0.551	0.691	0.100
C: Hardened (A+B)	−1.1	0.553	0.727	0.090
C: Hardened + DP (ε_{dp}=1)	−0.9	0.553	**0.522**	0.302

F.15 Hardened+DP at 14-Tier Scale

Table 21 reports the Hardened+DP (ε_{dp}=1) configuration on Iris, Wine, BreastCancer (3-fold × 3 seeds). To check that this DP-on-means hardening is not specific to those three datasets, we ran

the same configuration on the full 14-dataset numeric tier (3 folds, single seed). Table 22 reports per-dataset utility cost Δ_{real}, NN-MIA AUC, attribute Ridge R^2, and Wasserstein-1 distance. The 14-tier aggregate is consistent with the core-three result and shows that the hardening configuration scales beyond the small ablation, supporting the use of Table 21 as a sensitivity check.

Table 22: Hardened+DP (ε_{dp}=1) on the 14-dataset numeric tier (3 folds, single seed). Δ_{real} is the total accuracy cost of the hardened context against the real context, so it includes the synthesis cost, unlike Δ^{+}_{SDS} in Table 21. Mean utility cost is -2.8%, MIA stays near chance (mean 0.534), and mean attribute R^2 drops to 0.48 relative to the unhardened SDS baseline.

Dataset	N	d	C	Δ_{real}	**MIA**	**Attr** R^2	W_1
Iris	150	4	3	-2.00	0.598	0.552	0.412
Wine	178	13	3	-2.24	0.537	0.400	0.316
BreastCancer	569	30	2	-1.76	0.523	0.799	0.128
Diabetes	768	8	2	-4.30	0.536	0.209	0.093
Vehicle	846	18	4	-7.45	0.512	0.764	0.120
Segment	2000	19	7	-0.15	0.512	0.644	0.134
Banknote	1372	4	2	-0.73	0.500	0.521	0.046
Climate-Model	540	20	2	-7.78	0.527	-0.145	0.125
Ionosphere	351	34	2	$+1.14$	0.517	0.455	0.165
Seeds	210	7	3	-2.86	0.555	0.728	0.278
Dermatology	358	34	6	-1.40	0.550	0.462	0.458
Pendigits	1000	16	10	-3.00	0.529	0.612	0.176
Waveform	1000	40	3	-5.00	0.542	0.139	0.068
Optdigits	1000	64	10	-1.20	0.544	0.535	0.285
Mean (14)				$\mathbf{-2.77}$	**0.534**	**0.477**	**0.200**

F.16 CTGAN Comparison

CTGAN (Xu et al., 2019) under the same protocol (3-fold $\times$ 3-seed, $2\times$ oversampling, 300 epochs) incurs Δ=-48.9% at SDV defaults (Table 23). Even with the 600-epoch tuned config from Table 16 the mean stays around -20%.

Table 23: Ensemble vs. CTGAN. 3-fold $\times$ 3 seeds.

Dataset	**Acc-Real**	**Δ-Ens**	**Δ-CTGAN**
Iris	97.3	$\mathbf{-0.9}$	-20.7
Wine	98.9	$\mathbf{-0.6}$	-44.2
BreastCancer	97.5	$\mathbf{-1.2}$	-30.6
Multi3_Med	96.5	$\mathbf{-0.3}$	-59.7
Multi5_Med	92.9	$\mathbf{+1.4}$	-73.8
HighDim_50	98.4	$\mathbf{+0.5}$	-64.3
Mean	96.9	$\mathbf{-0.2}$	-48.9

F.17 PrivBayes Comparison

PrivBayes (Zhang et al., 2017) (Bayesian network under ε-DP, $\varepsilon \in \{1, 10\}$, same protocol) incurs $\Delta$$\approx$$-61\%$ at both privacy levels (Table 24). Discretization destroys the fine-grained structure that TabPFN uses.

F.18 AIM Comparison (Strongest DP-Tabular Baseline)

As the modern DP reference we benchmark AIM (McKenna et al., 2022), the strongest current marginal-based ε-DP tabular synthesizer. We had targeted the DP latent-diffusion model DP-TLDM (Zhu et al., 2024), but it exposes privacy only as an f-DP separation value (not a standard (ε, δ) budget) and has no public implementation. AIM instead offers a native (ε, δ) knob and state-of-the-art small-table utility, making it the more defensible DP comparator. We run it on the

Table 24: Ensemble vs. PrivBayes. 3-fold $\times$ 3 seeds. The Ensemble column is this run's own re-evaluation under its fold assignment and differs from the headline core-six value of -0.2% (Table 8).

Dataset	**Real**	**Δ-Ens**	**Δ-PB**(ε=1)	**Δ-PB**(ε=10)
Iris	97.3	$\mathbf{+0.2}$	-64.0	-63.8
Wine	98.5	$\mathbf{-0.7}$	-58.6	-59.2
BreastCancer	97.5	$\mathbf{-1.6}$	-45.4	-45.8
Multi3_Med	96.6	$\mathbf{-1.3}$	-63.5	-63.8
Multi5_Med	92.0	$\mathbf{-2.0}$	-72.0	-73.3
HighDim_50	96.8	$\mathbf{-0.5}$	-63.2	-63.2
Mean	96.5	$\mathbf{-1.0}$	-61.1	-61.5

core-six tier under a protocol parallel to PrivBayes (Table 24): 3-fold, $\varepsilon \in \{1, 10\}$, $\delta = 10^{-5}$, $n_{\text{bins}} = 32$, `max_model_size`= 80 (the authors' recommended setting), and `max_iters`= 200. Synthetic context size matches $|X_{\text{tr}}|$ (no oversampling). Table 25 reports the per-dataset utility gap $\Delta = \text{Acc}_T(\text{AIM synth}) - \text{Acc}_T(\text{real})$.

Table 25: AIM (McKenna et al., 2022) on the core-six tier (3-fold, $\delta = 10^{-5}$, `max_model_size`= 80, `max_iters`= 200). Δ is the AIM-synthetic-context vs. real-context TabPFN accuracy gap; closer to zero is better. SDS Δ on the same tier is -0.2% (Table 8).

Dataset	Δ AIM (ε=1)	Δ AIM (ε=10)
Iris	-64.0	-22.0
Wine	-53.9	-59.0
BreastCancer	-34.8	-25.1
Multi3_Med	-61.8	-49.5
Multi5_Med	-68.6	-69.6
HighDim_50	-56.8	-57.0
Mean (core six)	$\mathbf{-56.6}$	$\mathbf{-47.0}$

AIM at the recommended setting incurs a mean utility cost of -56.6% at $\varepsilon = 1$ and -47.0% at $\varepsilon = 10$ on the core six, better than PrivBayes (Table 24, $\approx -61\%$ at both privacy levels) but still substantially worse than the SDS Ensemble at -0.2% on the same tier (Table 8). The result reinforces the body's claim that DP-tabular synthesizers, including the strongest current marginal-based mechanism, lose the fine-grained signal that small-context TabPFN inference requires. SDS preserves utility while accepting that it does not satisfy a worst-case (ε, δ)-DP guarantee.

F.19 Gen-LRA and No-Box Density Membership Attacks

We audit SDS releases with Gen-LRA (Ward et al., 2025), the no-box local-likelihood-ratio attack designed to detect the local overfitting that near-chance global attacks can miss, alongside three density-ratio membership attacks (DOMIAS, a LOGAN-style synthetic-density score, and a discriminative classifier-ratio score). The protocol matches the DOMIAS audit (Table 51): on each of the 14 numeric-tier datasets (3 seeds) we split the data $50/50$ into members and a holdout, then split the holdout into a population reference R and a disjoint set of non-member test records. Member test records are a matched subset of the members. The SDS ensemble (α=2) is fit on the members and generates the release S. Gen-LRA follows the paper's Algorithm 1: a Gaussian KDE (Silverman bandwidth) is fit on R, and each test point $x^\star$ is scored by $\sum_{s \in S_k(x^\star)} \left[\log \hat{p}_{R \cup \{x^\star\}}(s) - \log \hat{p}_R(s)\right]$, the log-density gain of the synthetic release under adding $x^\star$ to the reference, summed over the k=10 nearest synthetic neighbors of $x^\star$. Higher means more member-like. All four attacks are scored so that higher is more member-like, and we report AUC in that natural direction (no per-attack sign flip, which would manufacture advantage from near-chance noise). This differs from the chance-centred advantage of Eq. (2) used elsewhere: here a below-chance AUC is reported as such rather than folded, because these attacks have a defined member-like direction. "Worst case" is the per-dataset maximum AUC over the four attacks, averaged over datasets.

As a positive control we replace the SDS release with a deliberately memorizing generator: noisy copies of the members (resampled with replacement, additive Gaussian noise at $0.25\times$ the per-feature standard deviation). A live attack must score well above chance there. Table 26 reports the 14-tier means. On SDS releases every attack, including Gen-LRA, stays near chance. On the memorizing generator the same attacks rise sharply, confirming that the audit is sensitive and that SDS's near-chance result reflects the release rather than a weak attack.

Table 26: No-box membership attacks on the 14-tier (3 seeds), AUC / chance-centered advantage / true-positive rate at 1% and 10% false-positive rate, means over datasets. On SDS releases all attacks stay near chance; on a deliberately memorizing generator (positive control) the same attacks fire, so the audit is live. "Worst case" is the per-dataset maximum AUC over the four attacks. The DOMIAS, LOGAN-density, and classifier-ratio rows are re-evaluated in this run and agree with the per-dataset means of Table 51 to within 0.005.

Attack	AUC	Adv	TPR@1%	TPR@10%
SDS release (the audited system)				
DOMIAS (van Breugel et al., 2023)	0.539	+0.039	0.022	0.122
LOGAN-density	0.533	+0.033	0.022	0.118
Classifier-ratio	0.538	+0.038	0.013	0.123
Gen-LRA (Ward et al., 2025)	0.544	+0.044	0.023	0.140
Worst case	**0.555**	+0.055	0.024	0.135
Memorizing generator (positive control)				
DOMIAS	0.701	+0.201	0.096	0.328
LOGAN-density	0.705	+0.205	0.124	0.317
Classifier-ratio	0.596	+0.096	0.019	0.162
Gen-LRA	0.690	+0.190	0.069	0.308
Worst case	**0.729**	+0.229	0.139	0.375

F.20 Random Forest Baseline (Per-Dataset)

The body assertion that "SDS+TabPFN matches or exceeds a 100-tree Random Forest on 10 of 14 datasets" (§6) is supported by a single, fully matched comparison. Table 27 reports real-context TabPFN, SDS-context TabPFN, and Random Forest (`RandomForestClassifier`, default depth, `n_estimators`=100, `n_jobs`=-1) on *identical* folds and identical size-capped data for all 14 datasets (5-fold $\times$ 3 seeds).

F.21 Wasserstein–Utility Empirical Diagnostic

To probe whether the quantity Theorem 1 singles out, the 1-Wasserstein distance $W_1(p_G, p^\star)$, tracks the observed utility gap on the actual datasets, we measured per-class assignment-based W_1 between the SDS-fitted generator and the empirical training measure on each of the 14 numeric-tier datasets (subsampled to 400 per class for tractability, mean over 3 resamplings, standardized features). We then correlated the per-dataset mean W_1 against $|\Delta_{\text{util}}|$ from Table 11.

Table 28 reports the per-dataset values. The cross-dataset correlation is indistinguishable from zero. This result is *consistent* with the body's framing of Theorem 1 as a sufficient distributional condition rather than a tight numerical bound, but does not strengthen it. The bound is $\mathbb{E}\|\Delta f\|_\infty \le L_f \cdot W_1$, and the empirical Lipschitz constant L_f varies by approximately two orders of magnitude across datasets (median 0.8 on Wine to 71.6 on Iris in Table 6, App. E). Cross-dataset numerical correlation of W_1 alone with utility is therefore not expected: the same W_1 implies very different prediction shifts on different datasets via different L_f. The within-dataset use of Theorem 1 as a design motivation (a smaller W_1 implies a smaller Lipschitz-multiplied prediction shift on a fixed dataset) remains valid. The cross-dataset numerical bound does not, and we state this to clarify that Theorem 1 is operationally useful for design but not predictively tight without dataset-specific Lipschitz calibration. A non-vacuous architectural Lipschitz upper bound for TabPFN (e.g., via certified-robustness tools such as `auto_LiRPA`) is the natural follow-up to close this gap. We leave it for future work.

Table 27: SDS+TabPFN vs. a 100-tree Random Forest on identical folds and size-capped data, all 14 datasets (5-fold $\times$ 3 seeds). This is an independent run with its own fold assignment, so its per-dataset SDS gaps differ from the headline Table 11 by up to 1.2 pp (e.g. Waveform -1.1 vs. -2.3). The headline numbers are those of Table 11. SDS matches or beats RF on 10/14, mean $+1.0$ pp. Three of the four losses are within noise (Diabetes, Climate-Model, Optdigits); Segment (-2.9 pp) also carries the second-largest synthesis gap on this tier, after Vehicle.

Dataset	N	d	**Acc-Real**	**Acc-SDS**	**SDS−RF (pp)**
Iris	150	4	0.971	0.969	+2.22
Wine	178	13	0.989	0.985	+0.76
BreastCancer	569	30	0.980	0.973	+1.40
Diabetes	768	8	0.760	0.760	−0.21
Vehicle	846	18	0.871	0.813	+6.78
Segment	2 000	19	0.990	0.946	−2.85
Banknote	1 372	4	1.000	0.997	+0.46
Climate-Model	540	20	0.909	0.911	−0.31
Ionosphere	351	34	0.945	0.938	+0.67
Seeds	210	7	0.964	0.957	+3.33
Dermatology	358	34	0.984	0.985	+1.21
Pendigits	1 000	16	0.985	0.972	+0.77
Waveform	1 000	40	0.866	0.855	+0.30
Optdigits	1 000	64	0.973	0.956	−0.53
Mean					**+1.00**

Table 28: Per-dataset assignment-based $W_1(p_G, p^\star)$ (mean over per-class W_1, 3 subsample repeats) vs. $|\Delta_{\mathrm{util}}|$ from Table 11. Pearson $r = 0.07$ ($p = 0.80$); Spearman $r = 0.23$ ($p = 0.43$). Neither is significant. N is the full dataset size, while $|\Delta_{\mathrm{util}}|$ comes from the size-capped protocol of Table 11.

Dataset	N	d	$W_1(p_G, p^\star)$	$\lvert\Delta_{\mathrm{util}}\rvert$ (%)
Iris	150	4	0.41	0.4
Wine	178	13	2.09	0.4
BreastCancer	569	30	2.82	1.2
Diabetes	768	8	1.46	0.4
Vehicle	846	18	2.00	5.7
Segment	2 310	18	0.90	4.4
Banknote	1 372	4	0.44	0.3
Climate-Model	540	20	3.98	0.1
Ionosphere	351	34	4.22	1.4
Seeds	210	7	0.66	0.2
Dermatology	358	34	3.45	0.0
Pendigits	10 992	16	1.35	1.2
Waveform	5 000	40	5.89	2.3
Optdigits	1 797	64	5.35	2.1

F.22 α=1 vs α=2

To match the equal-size regime of Proposition 2, we ran all 14 repeated-tier datasets at α=1 under the same 5-fold $\times$ 3-seed protocol. Mean Δ widens from -1.4% (α=2) to -1.8% (α=1). Oversampling helps on average, with Ionosphere as the lone exception.

F.23 Extended Obfuscation Audit (Non-Core)

On six non-core public benchmarks from the repeated tier (Table 30), MIA-Holdout averages 0.529. Segment (C=7) shows neural-discrimination AUC 0.778 but MIA-Holdout 0.509: distributional artifacts do not translate to record-level leakage.

Table 29: Δ (%) at α=1 vs. α=2. α=2 reproduces Table 11.

Dataset	$\Delta(\alpha{=}1)$	$\Delta(\alpha{=}2)$
Iris	-1.1	-0.4
Wine	-0.9	-0.4
BreastCancer	-1.3	-1.2
Diabetes	$+0.6$	$+0.4$
Vehicle	-6.9	-5.7
Segment	-4.6	-4.4
Banknote	-0.4	-0.3
Climate-Model	$+0.4$	$+0.1$
Ionosphere	-1.0	-1.4
Seeds	-1.3	-0.2
Dermatology	-0.6	-0.0
Pendigits	-2.3	-1.2
Waveform	-2.6	-2.3
Optdigits	-3.0	-2.1
Mean	$\mathbf{-1.8}$	$\mathbf{-1.4}$

Table 30: Obfuscation metrics on six non-core datasets, 5-fold $\times$ 3 seeds. **Log.** and **Neur.** are logistic and neural-net discriminator AUC (real vs. synthetic), **MIA** is the nearest-neighbor membership attack scored on a held-out fold, and W_1 is the marginal Wasserstein-1 distance, as in Table 14.

Dataset	N	d	**Log.**	**Neur.**	**MIA**	W_1
Diabetes	768	8	0.482	0.502	0.522	0.073
Vehicle	846	18	0.474	0.582	0.529	0.056
Segment*	2 000	19	0.484	0.778	0.509	0.045
Ionosphere	351	34	0.443	0.571	0.505	0.102
Seeds	210	7	0.485	0.480	0.547	0.066
Dermatology*	358	34	0.428	0.486	0.561	0.125
Mean			0.466	0.567	0.529	0.078

F.24 Near-Record Distance Audit

For each real training record x_i, we measure $\delta_i = \min_j \|x_i - x_j^s\|_2$ (features standardized). Distances are pooled across 25 runs per dataset. Table 31 reports percentiles, fractions within small radii, plus two baselines: R2R (leave-one-out nearest real-to-real) and Hld (median holdout-to-synthetic).

Table 31: Nearest-real-to-synthetic distances (standardized units). 25 runs per dataset.

Dataset	d	p_{50}	p_{95}	<0.25	<0.5	R2R p_{50}	Hld p_{50}
Iris	4	0.23	0.51	58.8%	94.4%	0.30	0.26
Wine	13	1.64	2.61	0.0%	0.0%	1.79	1.75
BreastCancer	30	2.12	4.16	0.0%	0.0%	2.13	2.17
Multi3_Med	20	2.17	2.68	0.0%	0.0%	2.33	2.22
Multi5_Med	25	2.48	2.96	0.0%	0.0%	2.65	2.56
HighDim_50	50	4.21	4.79	0.0%	0.0%	4.38	4.25

Holdout distances (Hld p_{50}) are similar to training distances, indicating no selective memorization.

F.25 DCR Release Threshold for Low-d Datasets

The body recommends a distance-to-closest-record (DCR) filter rule for the $d \lesssim 10$ regime, where Iris and similar low-d datasets have a non-trivial fraction of records within small radii. We evaluate a smooth, data-adaptive variant on six low-dimensional benchmarks (Iris, Banknote, Diabetes, Seeds, Blood-Transfusion, Phoneme, 3-fold $\times$ seed 42). For each fold, τ_{DCR} is set automatically as the 25th percentile of pairwise ℓ_2 distances within the standardized training set. This auto-scales to local data

density rather than fixing a global radius. A synthetic candidate with $\min_j \|x^s - x_j^{\text{train}}\| = \delta$ is then accepted with probability $\min(1, (\delta/\tau_{\text{DCR}})^2)$, repeating until each per-class quota of αN_c is filled or a candidate budget of 80 batches is exhausted.

The smooth ramp matters: a hard cutoff (accept iff $\delta \geq \tau_{\text{DCR}}$) eliminates the near-record fraction completely but raises NN-MIA advantage from 0.043 to 0.096 in our pilot, because the deterministic rule lets an adversary distinguish training records (always above τ_{DCR}) from holdout records (typical distance). The probabilistic version preserves the natural distribution of distances.

Table 32 reports the result. On the canonical Iris case (the 58.8% near-ε=0.25 flag in Table 31), NN-MIA halves from 0.117 to 0.053 at no accuracy cost. Seeds shows the cleanest gain: near-record drops $0.57 \to 0.26$ and NN-MIA $0.057 \to 0.042$ simultaneously. Blood-Transfusion is the dataset where the rule is least helpful: NN-MIA increases from 0.035 to 0.079, and the gap widens by 2.3 pp, but both remain within the leakage range reported for the single-release audits of Table 9 and Table 17. No fold hits the candidate budget in this setting (fail rate 0).

Table 32: DCR rule on six low-d datasets (τ_{DCR}=p_{25} within-train distance, smooth k=2 acceptance; 3-fold, seed 42). Acc gap is TabPFN(synth)−TabPFN(real); near-ε=0.5 is the fraction of training records with a synthetic neighbor within 0.5 standardized units after thresholding; NN-MIA is the adversarial advantage on train vs. holdout records scored by min-distance to the released context. The mean column is unweighted across the six datasets.

Metric	Iris	Bank.	Diab.	Seeds	Blood	Phon.
No threshold (baseline)						
acc gap (%)	+1.33	−0.58	+0.26	+0.00	+0.67	−6.00
near ε=0.5	0.94	0.99	0.03	0.57	0.97	0.83
NN-MIA adv	0.117	0.027	0.013	0.057	0.035	0.007
Smooth DCR (τ_{DCR}=p_{25}, k=2)						
acc gap (%)	+2.00	−0.22	−1.69	−1.43	−1.60	−4.60
near ε=0.5	0.86	0.96	0.005	0.26	0.95	0.56
NN-MIA adv	**0.053**	0.025	0.018	**0.042**	0.079	0.017

Cross-dataset means: baseline −0.72% / 0.72 / 0.043 vs. DCR −1.26% / 0.60 / **0.039**.

F.26 Scalability

Table 33 reports the Ensemble synthesizer on six representative larger datasets spanning $N \in [1{,}000, 5{,}000]$, $d \in [4, 61]$, and $C \in [2, 10]$. Table 34 shows how accuracy gap and geometry-attack R^2 vary as d grows from 10 to 100. $R^2 < 0.05$ throughout, and the NN-MIA AUC of 0.389 at d=100 is an advantage of 0.111 by Eq. (2), not a near-chance value. Because the index-paired metric scores a row-permuted copy of the real records the same way (App. L.10), we report these values for completeness and do not read them as evidence of protection.

Table 33: Ensemble on larger datasets. Context capped at 500.

Dataset	N	d	C	Real	Synth	Δ
Digits	1 797	61	10	96.8	94.3	−2.6
Synth (N1000, d20)	1 000	20	3	97.2	96.4	−0.8
Synth (N2000, d25)	2 000	25	4	95.3	95.6	+0.3
Synth (N3000, d30)	3 000	30	5	88.5	88.2	−0.4
Synth (N5000, d15)	5 000	15	3	97.4	97.9	+0.5
Banknote	1 372	4	2	100.0	99.3	−0.7
Mean				95.9	95.3	−0.6

G Neural Discriminator Architecture

The Neur. Disc. column of Table 14 is produced by the multilayer perceptron defined here. It takes a feature vector $x \in \mathbb{R}^d$ and returns the probability that x is a real-context record rather than a synthetic

Table 34: Dimensional sweep. $N = \max(500, 10d)$, $C = 3$.

d	N	**Real**	Δ	**MIA**	R^2
10	500	95.4	$+1.2$	0.477	0.014
20	500	96.8	-1.8	0.480	0.024
30	500	94.8	0.0	0.431	0.027
50	500	92.2	-0.2	0.427	0.045
75	750	90.8	-5.1	0.432	0.047
100	1 000	90.9	-4.3	0.389	0.048

one. The architecture is $\mathrm{MLP} : \mathbb{R}^d \to [0, 1]$ with hidden widths $(64, 32)$, ReLU activations, and a sigmoid output. Training uses cross-entropy loss with the Adam optimizer at learning rate 10^{-3} for 200 epochs and batch size 64. We report AUC via 5-fold cross-validation on a balanced set of X_{real} and X_{synth} rows.

H Broader Benchmark: 40 Datasets

Table 35 gives the per-dataset accuracy gap on a 40-dataset benchmark drawn from the OpenML CC18 suite (Bischl et al., 2019), the TabZilla hard-benchmark suite (McElfresh et al., 2023), and a small set of UCI classics. Each dataset is evaluated once, under the pipeline assigned by its categorical fraction (numeric-only Ensemble for continuous-dominated tables, Mixed-Type for categorical-heavy tables, marked *). No per-dataset selection on results was performed. The four datasets exceeding $|\Delta| > 9\%$ (Hill-Valley, Madelon, Wall-Robot, EEG-Eye-State, marked †) are high-dimensional or non-stationary scope-boundary cases.

I TabZilla-Hard Tier: Per-Dataset Results

Table 36 reports the per-dataset accuracy gap on the TabZilla hard-benchmark suite (McElfresh et al., 2023) (OpenML study 379). Of the 36 datasets, 31 are evaluated and 5 are excluded by the filters listed in App. M: audiology and lymph each have a class with fewer than six members, colic has $N < 100$, and albert and guillermo fail to load from OpenML. Each dataset is capped at $N{=}2000$ rows under 3-fold cross-validation, using the Mixed-Type pipeline of §4 at $\alpha{=}2$. The mean gap is -3.9%, and 24 of the 31 datasets stay within $|\Delta| \leq 5\%$. Real-context TabPFN is at chance on one-hundred-plants-texture (100 classes, 1.0% accuracy), so its $+0.0$ gap and its membership figure both describe a run in which the model learned nothing. Excluding it moves the tier mean to -4.0%. The mean NN-MIA AUC across the tier is 0.522. The highest value, 0.688, is that same degenerate dataset, and every other dataset sits at or below 0.554.

J Ablation Studies

J.1 Core-Six Synthesizer Ablation

Table 37 reports the per-dataset accuracy gap on the core six for each synthesizer configuration.

J.2 Oversampling Factor α

On BreastCancer ($N{=}569$, $d{=}30$), accuracy gap is stable across $\alpha \in \{0.5, 1, 2, 4, 8\}$ (within $\pm 0.2\%$, Table 38). MIA-AUC moves from 0.479 at $\alpha = 0.5$ to 0.444 at $\alpha = 4$, which by Eq. (2) is an increase in advantage from 0.021 to 0.056, so oversampling does not improve membership on this dataset.

J.3 Ensemble Weight w_{GMM}

On Wine ($N{=}178$, $d{=}13$), copula-only is the best single setting on both axes, closest to chance at MIA-AUC 0.464 (advantage 0.036 by Eq. (2)) with the smallest accuracy gap (-0.6%). The 50/50 default sits at 0.418 (advantage 0.082) and -0.7%, and the 75%-GMM setting is furthest from

Table 35: Full 40-dataset benchmark, each dataset evaluated once under the pipeline assigned by its categorical fraction (numeric-only Ensemble for continuous-dominated datasets, Mixed-Type for categorical-heavy datasets, marked *); † marks $|\Delta| > 9\%$. Mean gap -2.7%; 82.5% of datasets (33/40) within $|\Delta| \leq 5\%$.

Dataset	N	d	**Real**	Δ
Iris	150	4	98.0	$+0.0$
Wine	178	13	97.8	$+0.0$
BreastCancer	569	30	98.1	-1.2
Digits	1 797	61	97.6	-4.1
Multi3_Med	800	20	96.3	-0.9
Multi5_Med	1 000	25	91.5	-1.7
HighDim_50	2 000	50	97.1	-0.4
Vehicle	846	18	83.9	-6.4
Segment	2 310	18	97.3	-2.9
Diabetes	768	8	76.3	-1.7
SpamBase	4 601	57	93.9	-1.7
PC4	1 458	37	90.8	-1.4
KC1	2 109	21	85.4	-1.7
Banknote	1 372	4	100.0	-0.5
Heart-Statlog	270	13	77.0	$+7.4$
Blood-Transfus.	748	4	78.3	-0.1
Phoneme	5 000	5	83.5	-3.8
Australian	690	14	86.8	-1.9
Waveform	5 000	40	85.6	-2.9
Hill-Valley†	1 212	100	73.3	-12.2
Climate-Model	540	20	94.6	-0.9
Sonar	208	60	79.8	-3.4
Madelon†	2 600	500	68.7	-11.4
mfeat-factors	2 000	216	97.5	-1.2
mfeat-fourier	2 000	76	85.7	-3.5
Wall-Robot†	5 000	24	95.0	-9.3
Electricity	5 000	8	77.4	-2.7
EEG-Eye-State†	5 000	14	87.1	-9.9
Adult	5 000	14	84.2	-3.3
Credit-g	1 000	20	74.4	-2.9
Car*	1 728	6	96.5	-1.4
Nursery*	2 000	8	96.8	-0.5
Tic-Tac-Toe*	958	9	99.1	-8.2
Zoo	101	16	96.0	-2.0
Mushroom*	2 000	21	99.9	-1.2
CMC	1 473	9	55.6	$+0.5$
Anneal*	898	31	99.3	-3.8
Sick	3 772	27	97.8	-0.7
Bank-Marketing	5 000	16	89.9	$+0.3$
Chess-KR	3 196	36	97.9	-4.9
Mean (40)			89.0	-2.7

chance at 0.409 (advantage 0.091). The default is retained for its utility across the core six and the 14-tier (Tables 8 and 56), not on the evidence of this one dataset (Table 39).

J.4 Selection-vs-Random-Batch Leakage Audit

Algorithm 1 offers self-prediction context selection ($K{=}10$ candidates) as an option, while the headline runs sample a single context ($K{=}1$). Because selection is data-dependent (it queries TabPFN on real labels locally), one might worry that the selected batch leaks more than a single random batch. We measure this directly on the 14-tier at the headline-tier protocol of 5-fold $\times$ 3 seeds (seeds $\{42, 0, 1\}$, 15 runs per dataset and condition, $\alpha{=}2$): for each dataset, seed, and fold, fit the Ensemble synthesizer on X_{real}, then form two contexts. The $K{=}1$ context samples once at the seed, and the $K{=}10$ context runs the full self-prediction selection of §4 with $K{=}10$ candidates. We

Table 36: TabZilla-hard tier, all 31 evaluated datasets (3-fold, N capped at 2000, α=2). **Real** is real-context TabPFN accuracy in percent, Δ is SDS-context minus real-context accuracy, and NN-MIA is the nearest-neighbour membership AUC of Eq. (2), where chance is 0.5. The five excluded datasets are named in App. M. The mean row is computed from unrounded values.

Dataset	N	d	**Real**	Δ	**NN-MIA**
ada_agnostic	2000	48	85.5	−4.6	0.496
airlines	2000	7	59.2	+0.1	0.506
artificial-characters	2000	7	65.3	−9.0	0.551
Australian	690	14	87.4	−2.3	0.507
balance-scale	625	4	97.6	−4.2	0.529
Bioresponse	2000	1776	76.5	−4.9	0.507
cnae-9	1080	856	90.5	−6.6	0.511
credit-approval	666	15	86.8	−0.3	0.510
credit-g	1000	20	75.0	−1.5	0.526
electricity	2000	8	79.1	−2.7	0.522
elevators	2000	18	90.1	−4.8	0.508
GesturePhaseSegmentationProcessed	2000	32	60.0	−10.2	0.502
heart-h	270	13	81.8	+0.0	0.511
higgs	2000	28	70.5	−5.0	0.511
jasmine	2000	144	81.5	−3.3	0.510
jungle_chess_2pcs_raw_endgame_complete	2000	6	82.5	−3.3	0.513
kc1	2000	21	86.6	−16.5	0.503
mfeat-fourier	2000	76	89.0	−5.7	0.539
mfeat-zernike	2000	47	84.2	−0.8	0.525
MiniBooNE	2000	50	92.1	−3.4	0.515
monks-problems-2	601	6	99.5	−5.0	0.503
nomao	2000	118	94.8	−2.8	0.510
one-hundred-plants-texture	1599	64	1.0	+0.0	0.688
phoneme	2000	5	87.4	−5.6	0.511
poker-hand	2000	10	55.1	−4.6	0.505
profb	672	9	67.1	−1.0	0.489
qsar-biodeg	1055	41	87.9	−2.1	0.519
socmob	1156	5	93.5	−0.7	0.544
SpeedDating	1064	120	86.5	−3.0	0.554
splice	2000	60	96.5	−0.6	0.508
vehicle	846	18	85.1	−5.1	0.532
Mean (31)				**−3.9**	**0.522**

Table 37: Core six accuracy. Δ = Synth − Real (%). 5-fold × 5 seeds. Best $|\Delta|$ in **bold**.

	GMM-k2		**BIC-GMM**		**Copula**		**Ensemble**	
Dataset	Real	Δ	Real	Δ	Real	Δ	Real	Δ
Iris	97.1	−0.8	97.1	−0.4	97.1	**+0.1**	97.1	+0.2
Wine	98.9	−1.0	98.9	−0.2	98.9	−0.1	98.9	**−0.0**
BreastCancer	98.0	−2.1	98.0	−1.5	98.0	−1.3	98.0	**−1.1**
Multi3_Med	98.0	−1.5	98.0	**+0.4**	98.0	+0.6	98.0	+0.8
Multi5_Med	90.5	−3.9	90.5	+0.3	90.5	−0.2	90.5	**+0.1**
HighDim_50	92.4	−2.8	92.4	−1.8	92.4	−1.6	92.4	**−1.4**
Mean	95.8	−2.0	95.8	−0.5	95.8	−0.4	95.8	**−0.2**

then run identical leakage audits on both. Table 40 reports per-condition means across the 14 datasets and per-metric deltas.

The accuracy gap is nearly unchanged on this protocol: on the 14-tier the Ensemble is already strong, and selection moves the mean gap by 0.04 points. At 15 runs per cell the membership-advantage deltas Δ_{NN} and Δ_{QAPD} sit at the $\sim$0.001 scale, inside per-fold noise, and sign-favor K=10 (less leakage, not more), while the attribute and near-record deltas are flat. Selection can therefore be

Table 38: α ablation on BreastCancer. 3-fold $\times$ 3 seeds.

α	Acc-Synth	Δ	MIA-AUC
0.5	96.3	-1.2	0.479
1.0	96.5	-1.0	0.459
2.0	96.4	-1.1	0.447
4.0	96.4	-1.1	0.444
8.0	96.3	-1.2	0.453

Table 39: Ensemble weight ablation on Wine. 3-fold $\times$ 3 seeds.

w_{GMM}	Synth.	Acc	Δ	MIA
0.00	Copula	97.9	-0.6	0.464
0.25	25% GMM	97.4	-1.1	0.439
0.50	**50/50**	97.8	-0.7	0.418
0.75	75% GMM	97.2	-1.3	0.409
1.00	GMM only	96.8	-1.7	0.436

enabled without claiming an additional formal privacy guarantee. That choice rests on the empirical audit.

A proxy-selector variant replaces the TabPFN scorer with a local 100-tree Random Forest ranking the same 10 candidates: mean downstream TabPFN accuracy on the 14-tier is 0.930 for $K{=}1$, 0.929 for TabPFN-scored $K{=}10$, and 0.930 for RF-scored $K{=}10$, so selection does not require the production model. We therefore recommend the Random-Forest proxy as the default scorer, which removes the need to run the deployed model locally on real data, and we note that when the reader is small and open-weight the owner may simply run it in-house, in which case SDS is unnecessary. The headline numbers in this paper use no selection ($K{=}1$), and when selection is enabled the proxy matches the TabPFN scorer to within 0.001.

J.5 Repeated-Releases Composition Audit

The body's threat model considers a single SDS release. A natural follow-up question is whether membership and attribute leakage compound when the same data owner regenerates and transmits the synthetic context multiple times for the same dataset, e.g. across model updates or repeated re-scoring. We test this directly: fix the real training set $\mathcal{D} = (X, y)$, fit the Ensemble synthesizer on a member half, and for $k \in \{1, 2, 5, 10, 20\}$ form a pool $\widetilde{\mathcal{D}}_k = \bigcup_{r=1}^{k} (\widetilde{X}^{(r)}, \widetilde{y}^{(r)})$ of k successive releases (different sampling seed per release, pool size grows linearly to $k\alpha N$). We then run the class-conditional NN-MIA of Stadler et al. (2022) and a Ridge attribute-inference attack against the pool. Table 41 reports the result on six representative 14-tier datasets spanning N from 75 to 5500 training members.

The largest dataset, Pendigits ($N{=}5496$ members, pool size 219,840 at $k{=}20$), is rock-stable at AUC 0.522–0.525 across all k, providing the strongest evidence at scale that pooling does not amplify membership leakage. Table 60 re-runs the sweep with the member set capped at 500 records, where Pendigits drifts from 0.553 to 0.595 over 200 releases, so the stability reported here is a property of the large member set rather than of the dataset. The result complements the $K{=}1$ vs. $K{=}10$ selection audit of Table 40: that audit shows a single release is robust to data-dependent selection. The present audit shows that many such releases are robust to longitudinal pooling.

J.6 DCR-Band Mitigations for Small-N Compounding

The Iris result of Table 41 ($+0.11$ NN-MIA growth over 20 releases) raises a natural deployment question: can a stricter version of the optional DCR filter (Algorithm 1, line 14) suppress longitudinal compounding without sacrificing utility? Concretely, we test two filter families against the unfiltered baseline (the base column reports the unfiltered advantage as re-measured within this audit's own runs, so it differs slightly from Table 41):

Table 40: Selection (K=10) vs. random batch (K=1) on the 14 repeated-tier datasets at the headline-tier protocol (5-fold $\times$ 3 seeds, 15 runs per cell). This is a separate run, so its mean gap (-0.99) differs from the headline -1.4 of Table 11. The comparison of interest is between the two rows, not against the headline. NN-MIA and QAPD-MIA are reported as adversarial advantage $\max(\text{AUC}, 1-\text{AUC}) - 0.5$. Attr-$R^2$ excludes Ionosphere and Optdigits (degenerate Ridge on both conditions). Near-record rate is the fraction of training records with a synthetic neighbor within ε=0.5 standardized units. Every leakage axis is flat or sign-favors K=10 (less leakage), confirming that self-prediction selection adds no measurable leakage cost.

Setting	**Acc gap (%)**	**NN-MIA adv**	**QAPD adv**	**Attr-R^2**	**Near-rec.** (ε=0.5)
K=1 random batch	-0.95	0.044	0.080	0.573	0.228
K=10 selected batch	-0.99	0.043	0.079	0.573	0.228
Δ (K=10 $-$ K=1)	-0.04	-0.001	-0.001	$\approx$0	$\approx$0

Table 41: Repeated-releases composition: NN-MIA AUC and attribute R^2 as a function of the number of successive SDS releases $k \in \{1, 2, 5, 10, 20\}$ pooled by an attacker. NN-MIA AUC stays within ± 0.03 of the k=1 value on BreastCancer, Diabetes, Vehicle, and Pendigits; attribute R^2 moves by at most $+0.03$ (Wine). Growth concentrates on the two smallest member sets: Iris (training half N=75, $+0.11$ over 20 releases) and, mildly, Wine ($+0.06$ at its k=5 peak), reflecting the well-known sample-size sensitivity of class-conditional NN-MIA on tiny datasets. Longitudinal release leaves leakage flat on the four larger datasets and raises it on the two smallest.

Dataset	k=1	k=2	k=5	k=10	k=20	**metric**
Iris	0.602	0.613	0.656	0.680	0.715	NN-MIA AUC $\downarrow$
Wine	0.648	0.655	0.703	0.688	0.672	NN-MIA AUC $\downarrow$
BreastCancer	0.503	0.515	0.512	0.510	0.514	NN-MIA AUC $\downarrow$
Diabetes	0.563	0.566	0.574	0.561	0.569	NN-MIA AUC $\downarrow$
Vehicle	0.547	0.527	0.522	0.534	0.537	NN-MIA AUC $\downarrow$
Pendigits	0.525	0.523	0.523	0.522	0.523	NN-MIA AUC $\downarrow$
Iris	0.889	0.889	0.891	0.891	0.892	attr R^2 $\downarrow$
Wine	0.620	0.626	0.649	0.649	0.650	attr R^2 $\downarrow$
BreastCancer	0.999	0.999	0.999	0.999	0.999	attr R^2 $\downarrow$
Diabetes	0.106	0.117	0.119	0.119	0.119	attr R^2 $\downarrow$
Vehicle	0.790	0.799	0.799	0.798	0.798	attr R^2 $\downarrow$
Pendigits	0.649	0.648	0.649	0.650	0.651	attr R^2 $\downarrow$

- **Single-threshold** (the original DCR rule): retain candidates with $\delta(x) \geq \tau_{\text{DCR}}$ for $\tau_{\text{DCR}} \in \{0.5, 0.8, 1.0, 1.5\}$ standardized units.
- **Two-sided band-pass**: retain candidates with $\delta(x)$ in a percentile band $[p_{\text{lo}}, p_{\text{hi}}]$ of the within-batch DCR distribution (scale-invariant), or in an absolute band $[\tau_{\min}, \tau_{\max}]$. The motivation is that single-threshold filtering creates an asymmetric synth distribution (synth biased away from members but not away from holdouts), which can *invert* the NN-MIA score sign and preserve adversarial advantage with reversed direction. A two-sided band keeps synth in the typical-distance band of natural training-distribution geometry.

Table 42 reports chance-centered advantage $|\text{AUC} - 0.5|$ at k=20 across eight datasets and seven filter settings.

Finding 1 (the band-pass works for very-small-N compounding cases). The asymmetric percentile band $(p_{\text{lo}}, p_{\text{hi}})$=$(10, 75)$ cleanly suppresses Iris compounding ($0.161 \rightarrow 0.006$, near-perfect chance recovery, below the 14-tier mean) and substantially reduces Wine ($0.202 \rightarrow 0.065$). Both correspond to N_{mem}<100 where BIC selects k=1 component per class.

Finding 2 (no filter setting universally dominates baseline). On six larger datasets ($N_{\text{mem}} \in [284, 5496]$: BreastCancer, Diabetes, Vehicle, Banknote, Segment, Pendigits) where the unfiltered NN-MIA already stays at chance, every percentile band introduces score-sign asymmetry that pushes the chance-centered advantage *above* baseline. The interaction is not predictable from $\{N, d, C\}$

alone: e.g., the absolute band $(0.5, 1.5)$ matches baseline on Pendigits (0.021 vs 0.023) and improves Diabetes ($0.062 \to 0.031$), but distorts Banknote ($0.031 \to 0.198$).

Table 42: Chance-centered NN-MIA advantage $|\mathrm{AUC}-0.5|$ at k=20 pooled releases, across eight datasets and seven filter settings (five percentile bands plus two absolute bands). $\mathbf{p}_{\mathrm{lo}}$–$\mathbf{p}_{\mathrm{hi}}$ denotes a percentile band on within-batch DCR; $\mathbf{abs}(\mathbf{a}, \mathbf{b})$ denotes an absolute band $\delta(x) \in [a, b]$ in standardized units. Bold = best filter for that dataset (lowest advantage). The two-sided $(10, 75)$ band wins on small-N compounding cases (Iris, Wine) but no filter dominates baseline on larger datasets.

Dataset (N_{mem}, d, C)	**base**	**p10-90**	**p25-75**	**p33-67**	**p10-75**	**p25-90**	**abs(0.5,1.5)**	**abs(0.8,1.5)**
Iris $(75, 4, 3)$	0.161	0.028	0.231	0.285	**0.006**	0.283	0.228	0.066
Wine $(89, 13, 3)$	0.202	0.086	0.083	0.172	**0.065**	0.089	0.249	0.247
BreastCancer $(284, 30, 2)$	**0.028**	0.132	0.255	0.288	0.126	0.246	0.064	0.064
Diabetes $(384, 8, 2)$	0.062	0.104	0.237	0.277	0.099	0.226	**0.031**	0.154
Vehicle $(423, 18, 4)$	**0.050**	0.163	0.289	0.338	0.170	0.267	0.088	0.057
Banknote $(686, 4, 2)$	**0.031**	0.217	0.337	0.368	0.227	0.313	0.198	0.062
Segment $(1155, 18, 7)$	**0.048**	0.084	0.174	0.184	0.104	0.157	0.168	0.119
Pendigits $(5496, 16, 10)$	0.023	0.082	0.173	0.217	0.086	0.168	**0.021**	0.200

Why no universal rule exists. The asymmetry of any DCR filter is fundamental to the locally-deployable threat model: the data owner can compute $\delta(x)$ only against members, not against holdouts (which the owner does not have). Filtering against members alone biases the synth distribution *relative to members specifically*, distorting the symmetry that the unfiltered distribution preserves.

Deployment rule. We therefore recommend the band-pass DCR filter $(p_{\mathrm{lo}}, p_{\mathrm{hi}})=(10, 75)$ *only* when the data owner satisfies a small-N structural diagnostic: $N_{\mathrm{mem}}<100$. On the two datasets in that regime the BIC-selected GMM also has k=1 component on every class, though the two conditions are not equivalent in general. Both are locally checkable. For larger N_{mem}, the unfiltered audit of Table 41 already stays at chance under repeated releases, and additional filtering is not recommended.

J.7 Per-Feature Attribute Inference

The per-feature breakdown (Table 43) shows a structural pattern: strongly correlated features (e.g., BreastCancer radius and perimeter, $R^2>0.99$) are nearly fully recoverable, while weakly correlated ones (e.g., Wine color intensity, R^2=0.09) are not. This reflects the utility-leakage coupling of Theorem 3: a generator that preserves the correlations a TabPFN-style model exploits also hands them to an attacker with access to the synthetic context.

Table 43: Attribute inference R^2 (ridge on X_{synth}, eval on X_{real}), per-dataset synthesizer configuration of Table 37, no percentile clipping. The real-data-only subset (Iris, Wine, BreastCancer) averages R^2=0.69; the 0.77 figure quoted in §6 and Table 7 is the same real-data core measured under the hardening protocol (Ensemble synthesizer, $[1, 99]$-percentile clip, Table 21), which is the number we recommend citing.

Dataset	d	**Mean**	**Max**	**Min**	>0.9	>0.8	<0.5
Iris	4	0.80	0.97	0.46	2/4	3/4	1/4
Wine	13	0.47	0.81	0.09	0/13	1/13	8/13
BreastCancer	30	0.81	1.00	0.40	14/30	17/30	1/30
Multi3_Med	20	0.68	1.00	−0.10	14/20	14/20	6/20
Multi5_Med	25	0.83	1.00	−0.06	21/25	21/25	4/25
HighDim_50	50	0.79	1.00	−0.07	40/50	40/50	10/50
Mean		0.73					

J.8 Nonlinear Attribute-Inference Attackers

The attribute-inference attacker in §6 is Ridge. Because Ridge is linear, one might suspect that it understates leakage on a generator that preserves nonlinear or class-conditional structure (the GMM/Copula ensemble does both). We re-run the attribute-inference protocol with three nonlinear

attackers (Random Forest with 100 trees of depth 20, Extra Trees with 100 trees of depth 20, and HistGradientBoosting with 200 iterations of depth 6) on the core six and the 14-tier numeric audit, with the same $[1, 99]$-percentile clip on X_{synth} used in the broad audit.

Table 44: Mean attribute-inference R^2 across attackers, on the core six (Table 37 synthesizers) and the 14-tier numeric audit. Synthesizers are released as-is; per-column R^2 are clipped at -1 for averaging (a degenerate Ridge on Ionosphere/Optdigits would otherwise dominate). Δ is the attacker's mean minus Ridge's mean. Nonlinear attackers gain at most $+0.010$ over Ridge on the 14-tier and underperform Ridge by 0.09–0.14 on the core six (which is mostly synthetic, so Ridge happens to be near-optimal). The 14-tier column is the same run reported against a population control in Table 62.

Attacker	**Core-six** R^2	Δ vs. Ridge	**14-tier** R^2	Δ vs. Ridge
Ridge	0.79	–	0.60	–
ExtraTrees	0.66	-0.13	0.61	$+0.010$
RandomForest	0.65	-0.14	0.61	$+0.004$
HistGradientBoosting	0.70	-0.09	0.58	-0.026

Table 44 shows that Ridge is a fair (and, on synthetic-feature core-six datasets, slightly conservative) estimate of attribute leakage. Per-dataset, the only places nonlinear attackers materially exceed Ridge are low-dimensional ones with strongly nonlinear feature structure (Banknote d=4: Ridge 0.53, RF 0.76, and Pendigits d=16: Ridge 0.77, ExtraTrees 0.84), and the tier mean still moves by ≤ 0.03. We therefore retain Ridge as the headline attribute-inference attacker in §6, including the hardening ablation.

K Imbalanced-Binary Utility-Metric Audit

The headline utility metric in §2 is class-balanced top-1 accuracy when the test set T has imbalanced classes. To check whether this single metric understates utility loss on heavily imbalanced datasets, we audit balanced accuracy, minority-class precision, recall, and F1, and PR-AUC, the area under the precision-recall curve, on seven binary OpenML datasets with minority-class rate $\leq 35\%$: kc1 (15.5% minority), ada_agnostic (24.8%), blood-transfusion-service-center (23.8%), MiniBooNE (28.1%), phoneme (29.3%), credit-g (30%), and diabetes (34.9%). Protocol: 5-fold $\times$ 2-seed $= 10$ runs per dataset, N capped at 1500 to keep TabPFN context size in range. The profb dataset was excluded because the OpenML version we loaded contains NaN entries that the synthesizer cannot handle natively. We did not impute.

Tuning heuristic. The default $\alpha = 2$ is well-chosen on six of the seven datasets here. On severely imbalanced data (minority rate $\leq 20\%$), $\alpha = 1$ may be preferable. We treat this as a preliminary heuristic rather than a recommendation, since kc1 (15.5% minority) is the only dataset in that regime in our suite. The full α-sweep that supports this rule is in Table 46, and Table 47 verifies that the choice does not relax privacy.

Table 45 reports the SDS-vs-real-context delta on each metric. We split out minority-class precision and recall (the precision and recall computed against the minority label, since minority recall is the operationally relevant detection rate on imbalanced datasets) to surface the precision–recall tradeoff that PR-AUC compresses into a single number.

Three observations follow. First, balanced accuracy (our headline metric) is the most stable across the suite. On three datasets (kc1, ada_agnostic, diabetes) the synthetic context outperforms the real context, suggesting the SDS oversampling factor α=2 partially counteracts the minority-class disadvantage that real-context TabPFN faces with small N_c. Second, the precision–recall split shows that on four of seven datasets, the synthetic context shifts the operating point toward higher minority recall at the cost of lower minority precision (mean -10.3%). Minority F1 averages -2.8% but is positive on three datasets where the recall gain outweighs the precision drop. Third, PR-AUC drops most. Because PR-AUC depends only on the ranking of predicted scores, a monotone calibration change cannot explain it, so the synthetic context is reordering minority examples rather than rescaling them. This matters for threshold-sensitive applications such as screening pipelines with explicit

Table 45: Utility-metric audit on seven imbalanced binary datasets (SDS synthetic-context → real-context delta per metric). Precision, recall, and F1 are reported on the minority class. Balanced accuracy (the headline metric of §2) moves by -0.7% on average and *improves* on three datasets (bold). The synthetic context shifts the precision–recall operating point toward higher minority recall and lower minority precision on most datasets, leaving F1 small in magnitude while PR-AUC still falls, which reflects a change in how minority examples are ranked rather than calibration alone.

Dataset	min%	Bal-Acc	Prec. min	Recall min	F1 min	PR-AUC
kc1	15.5%	$+\mathbf{3.7}\%$	-29.2%	$+\mathbf{19.7}\%$	$+\mathbf{3.9}\%$	-14.6%
ada_agnostic	24.8%	$+\mathbf{2.2}\%$	-12.6%	$+\mathbf{12.5}\%$	$+\mathbf{0.6}\%$	-4.7%
blood-transfusion	23.8%	-3.9%	-0.2%	-9.3%	-10.2%	-1.7%
MiniBooNE	28.1%	-0.8%	-10.1%	$+\mathbf{3.6}\%$	-3.6%	-3.2%
phoneme	29.3%	-5.8%	-16.7%	-1.7%	-9.5%	-11.9%
credit-g	30.0%	-0.7%	-1.9%	-0.8%	-1.4%	-4.7%
diabetes	34.9%	$+\mathbf{0.4}\%$	-1.4%	$+\mathbf{2.6}\%$	$+\mathbf{0.7}\%$	-1.7%
Mean		-0.7%	-10.3%	$+\mathbf{3.8}\%$	-2.8%	-6.1%

operating points, and is a residual cost of synthetic-context release that the headline accuracy metric does not surface.

The phoneme dataset is the largest outlier, with most metrics dropping by 5%–17%. It has one of the smallest feature dimensions of the suite (d=5, behind blood-transfusion at d=4) and the largest sample size at the cap, so the BIC-GMM ensemble's per-feature density estimates are spread thin. Excluding phoneme, the balanced-accuracy gap is $+0.15\%$ mean across the remaining six datasets, negligible and well inside the headline-tier -1.4% of §6.

α-sweep on the four most-affected datasets. Table 46 reports the SDS→real delta for each utility metric across $\alpha \in \{0.5, 1, 2, 4\}$ on the four datasets with the largest precision drop at the default (5-fold × 1-seed). The sweep supports the heuristic. (i) On kc1 (15.5% minority) the default $\alpha = 2$ trails $\alpha = 1$ on minority precision, F1, and PR-AUC: dropping to $\alpha = 1$ improves minority F1 from $+3.3\%$ to $+\mathbf{5.2}\%$ and PR-AUC from -13.7% to $-\mathbf{9.8}\%$, with the precision drop also smaller (-20.6% vs. -31.0%). (ii) On the other three datasets the default is at or near the per-dataset best, and tuning produces sub-percent gains. The phoneme outlier improves only at $\alpha = 4$ and stays the worst dataset at every α, consistent with the underlying generator-quality cause.

Table 46: α-sweep on the four imbalanced binary datasets with the largest precision drop at α=2. Entries are SDS→real deltas in percentage points (5-fold × 1-seed). Bold values are the best per column within each dataset block, and the bold α marks the tuning rule's choice, α=1 for kc1 ($\leq 20\%$ minority) and α=2 otherwise.

Dataset	α	Bal-Acc	Prec. min	Recall min	F1 min	PR-AUC
kc1 (15.5%)	0.5	-1.7%	-28.6%	$+0.9\%$	-4.5%	-14.1%
	1.0	$+3.3\%$	$-\mathbf{20.6}\%$	$+11.0\%$	$+5.2\%$	$-\mathbf{9.8}\%$
	2.0	$+3.5\%$	-31.0%	$+18.9\%$	$+3.3\%$	-13.7%
	4.0	$+\mathbf{6.8}\%$	-36.7%	$+\mathbf{32.8}\%$	$+\mathbf{7.9}\%$	-16.7%
ada_agnostic (24.8%)	0.5	0.0%	$-\mathbf{10.6}\%$	$+5.1\%$	-1.8%	-4.0%
	1.0	$+0.1\%$	-11.9%	$+6.4\%$	-1.9%	-5.4%
	2.0	$+\mathbf{1.6}\%$	-14.4%	$+\mathbf{11.7}\%$	$-\mathbf{0.7}\%$	-5.1%
	4.0	$+1.2\%$	-15.0%	$+11.2\%$	-1.3%	$-\mathbf{3.8}\%$
MiniBooNE (28.1%)	0.5	-2.4%	-12.6%	$+1.5\%$	-5.8%	-4.3%
	1.0	-1.8%	-9.9%	$+1.2\%$	-4.5%	-3.3%
	2.0	$-\mathbf{0.7}\%$	$-\mathbf{9.2}\%$	$+\mathbf{3.2}\%$	$-\mathbf{3.2}\%$	-3.4%
	4.0	-1.4%	-11.3%	$+3.0\%$	-4.4%	$-\mathbf{3.2}\%$
phoneme (29.3%)	0.5	-8.1%	-18.8%	-5.8%	-12.5%	-17.2%
	1.0	-4.8%	$-\mathbf{15.4}\%$	-0.5%	-8.4%	-15.9%
	2.0	-5.8%	-17.5%	-0.9%	-9.7%	-13.6%
	4.0	$-\mathbf{3.5}\%$	-16.3%	$+\mathbf{3.9}\%$	$-\mathbf{7.1}\%$	$-\mathbf{11.3}\%$

Privacy under the heuristic. To verify that the $\alpha = 1$ choice on severely imbalanced data does not relax privacy, we re-run four leakage audits at $\alpha = 1$ versus $\alpha = 2$ on all seven datasets (Table 47): nearest-neighbor MIA AUC (Eq. (2)), attribute-Ridge R^2 (Eq. (4)), near-record fraction at $\varepsilon = 0.5$ standardized units, and a logistic-discriminator AUC (real vs. synthetic, 5-fold CV). Across the three direct privacy metrics (NN-MIA, attribute R^2, near-record fraction, 21 dataset-metric pairs), $\alpha = 1$ is at or below $\alpha = 2$ on $19/21$, and the two cases where $\alpha = 2$ is lower are by margins ≤ 0.006, well within per-fold noise. Mean NN-MIA AUC is 0.507 at $\alpha = 1$ versus 0.510 at $\alpha = 2$ (both near the chance level 0.5). Mean attribute R^2 is 0.424 versus 0.431. Mean near-record fraction is 0.295 versus 0.331. The heuristic therefore tightens privacy on most metrics while improving utility on the targeted minority-class regime.

Table 47: Privacy/leakage audit at α=1 vs. α=2 on the seven imbalanced binary datasets (5-fold $\times$ 1-seed). NN-MIA AUC (Eq. (2)) is closer to chance (0.5) on $5/7$ datasets at α=1; attribute R^2 (Eq. (4)) is at or below α=2 on $7/7$; near-record fraction at ε=0.5 is at or below on $7/7$. The logistic discriminator shows no above-chance signal at either α (all values ≤ 0.5), and its below-chance values, which arise from cross-validated training, are not folded here. Boldface marks the privacy-better setting for each (dataset, metric).

		NN-MIA		**Attr.** R^2		**Near**(ε=.5)		**Log-disc**	
Dataset	**min%**	α=1	α=2	α=1	α=2	α=1	α=2	α=1	α=2
kc1	15.5%	**0.502**	0.504	0.865	0.865	**0.471**	0.511	0.472	0.487
ada_agnostic	24.8%	0.505	**0.499**	**0.193**	0.214	**0.003**	0.008	0.458	0.454
blood-transfusion	23.8%	**0.503**	0.515	0.666	0.666	**0.955**	0.979	0.320	0.309
MiniBooNE	28.1%	0.512	**0.509**	**0.746**	0.752	0.000	0.000	0.464	0.439
phoneme	29.3%	**0.507**	0.512	**0.089**	0.093	**0.632**	0.804	0.487	0.479
credit-g	30.0%	**0.510**	0.519	**0.151**	0.163	0.000	0.000	0.457	0.463
diabetes	34.9%	**0.509**	0.515	**0.261**	0.265	**0.007**	0.014	0.473	0.474
Mean		**0.507**	0.510	**0.424**	0.431	**0.295**	0.331	0.447	0.443

L Additional Robustness and Generalization Experiments

L.1 Class- and Confidence-Stratified Selection Audit

Table 40 shows K=10 vs. K=1 selection is flat in aggregate on the 14-tier. To rule out a subgroup effect hidden by that aggregate, we repeat the audit stratified by minority class (the least-represented class per dataset) and by real-context self-prediction confidence tercile (bottom tercile = decision-boundary records), same protocol (5-fold $\times$ 3 seeds, 15 runs/dataset).

Table 48: Selection audit stratified by minority class and confidence tercile, means over the 14-dataset tier. $\Delta = K$=10 $- K$=1; negative NN-MIA Δ favors K=10.

Stratum / metric	**K=1**	**K=10**	Δ	**max** $\lvert\Delta\rvert$
Minority NN-MIA AUC	0.568	0.563	-0.005	0.033
Minority nearest-synth. dist.	2.43	2.44	$+0.007$	0.14
Minority coverage ratio	1.004	1.003	-0.001	0.13
Minority test recall	0.859	0.862	$+0.003$	0.034
Low-conf. tercile NN-MIA AUC	0.553	0.552	-0.001	0.023
Low-conf. nearest-synth. dist.	2.18	2.17	-0.003	0.028
High-conf. tercile NN-MIA AUC	0.475	0.480	$+0.005$	0.048

Every stratified delta in Table 48 sits inside per-fold noise. The largest per-dataset membership movements in the minority and low-confidence strata (Wine minority MIA -0.033, Seeds low-confidence MIA -0.023) favor K=10. Minority coverage stays at parity under both conditions, and minority recall is marginally better under selection. This is consistent with the selection mechanism itself: it picks among whole candidate batches drawn i.i.d. from the same fitted generator, with no direct channel to concentrate synthetic mass near any specific record or class. The maximum

information selection could in principle encode about the training data is bounded by $\log_2 K$, about 3.3 bits per release for $K{=}10$, spread across the entire released context.

L.2 Heterogeneous Repeated Releases

Table 41 pools k homogeneous releases from one fitted generator. Here we pool 18 *heterogeneous* releases per dataset (Table 49), crossing all three generators (Ensemble, BIC-GMM, Copula), $\alpha \in \{1, 2, 3\}$, and two independent generator refits, on all 14 numeric-tier datasets (member sets 75–1,000 records, pools up to 36,000 rows).

Table 49: Heterogeneous pooled-release audit, $r{=}1$ (single release) to $r{=}18$ (all pooled). **Pooled err.** is the deviation of the attacker's pooled per-class mean/correlation estimate from the real-data statistics (shown $r{=}1{\to}18$); **Gen. err.** is the fitted generator's own sampling error at $r{=}18$, shown for reference.

Dataset	NN-MIA	Attr-R^2	Geom-R^2	Pooled err.	Gen. err.
Iris	0.568→0.655	0.81→0.82	0.005→0.031	0.017→0.003	0.005
Wine	0.635→0.690	0.53→0.59	0.048→0.063	0.035→0.006	0.007
BreastCancer	0.509→0.518	0.93→0.94	0.015→0.015	0.009→0.004	0.007
Diabetes	0.563→0.557	0.25→0.26	0.013→0.008	0.020→0.006	0.008
Vehicle	0.496→0.556	0.84→0.86	0.007→0.010	0.019→0.004	0.002
Segment	0.535→0.550	0.58→0.59	0.005→0.005	0.030→0.007	0.008
Banknote	0.525→0.539	0.55→0.55	0.004→0.003	0.094→0.011	0.020
Climate-Model	0.514→0.574	0.06→0.13	0.031→0.032	0.026→0.006	0.006
Ionosphere	0.504→0.519	0.50→0.66	0.049→0.038	0.193→0.041	0.047
Seeds	0.601→0.650	0.86→0.87	0.022→0.014	0.017→0.003	0.002
Dermatology	0.618→0.619	0.55→0.63	0.030→0.035	0.043→0.015	0.017
Pendigits	0.555→0.599	0.76→0.77	0.011→0.011	0.036→0.009	0.012
Waveform	0.576→0.610	0.24→0.30	0.015→0.016	0.057→0.017	0.013
Optdigits	0.566→0.598	0.51→0.56	0.017→0.016	0.063→0.017	0.013

Pooling converges toward the generator, not the data: the attacker's pooled per-class mean/correlation estimate tracks the generator's own sampling statistics (last column, matched to within 0.01 on every dataset at $r{=}18$), so pooling converges to the generator's estimate of the data rather than to the records themselves. The index-paired geometry score stays below 0.063 throughout, which we report for completeness rather than as evidence, since that metric does not separate a synthetic release from a row-permuted copy of the real records (App. L.10). NN-MIA compounds by 0.05–0.09 on five datasets (Iris, Wine, Vehicle, Climate-Model, Seeds) and by less than 0.045 on the other nine, consistent with the homogeneous-release audit (App. J.5). Pooling out to 200 releases moves NN-MIA by at most $+0.075$ on twelve of the fourteen datasets (App. L.15).

L.3 Error Redistribution by Prediction Margin

App. J.8 and Table 45 address whether SDS's utility cost concentrates in specific classes. Here we stratify by prediction confidence directly. On the 14-tier (5-fold $\times$ 3 seeds, identical folds for real and SDS context), we split each test set into four quartiles by real-context prediction margin (top-1 minus top-2 class probability) and measure the extra error SDS introduces in each quartile.

Only 4.4% of predictions change between contexts at all. In the highest-margin quartile the extra error is $+0.0003$ on average (max $+0.004$ across all 14 datasets). Confident predictions are nearly untouched. In the lowest-margin quartile it is $+0.028$ on average, concentrated on Segment ($+0.17$) and Vehicle ($+0.10$), the two datasets with the largest overall accuracy gaps. Mean worst-class recall drop is 0.044, exceeding 0.10 only on Segment (0.20) and Vehicle (0.11). Every other dataset's per-class recall moves by less than 0.08. The synthetic-context cost is therefore a low-margin effect that explains the headline gap rather than hiding a separate one, consistent with the minority-class shift already reported in Table 45.

A follow-up generator-tuning study on the two worst datasets separates their causes. Vehicle's gap is a uniform boundary blur across all four classes (per-class recall drops of 0.05–0.07 with well-matched class moments) and responds to generator capacity and sample budget: a GMM-only

synthesizer with $\max k{=}16$ components and $\alpha{=}4$ oversampling narrows the gap from -7.0% to -4.3%, and $\alpha{=}8$ alone reaches -5.0%. Segment's gap is different in kind: it is confined to a single class (recall $0.97 \to 0.79$, all six other classes move by at most 0.02), and no tested configuration moves it (oversampling, more components, tighter covariance regularization, and best-of-20 candidate selection all leave -4.0% to -4.3%, dropping the copula makes it worse, -6.1%), indicating higher-order class structure outside the GMM-plus-copula family rather than an estimation or sample-size problem. Practitioners hitting a Vehicle-like gap can trade release size for accuracy. A Segment-like single-class gap calls for a richer generator for that class or accepting the documented cost.

L.4 Covariance Stability Under Subsampling

App. C.15 gives a conditional local-stability diagnostic for the copula's dependence estimator. We audit it directly: for every class in every 14-tier dataset, we take 100 resamples at 90% of the class size (subsampling, gentler than full bootstrap) and refit the exact copula estimator (Ledoit–Wolf shrunk covariance on rank-normal scores when $N_c < 3d$), comparing pairwise correlations, the top-3 eigenspace, and BIC-selected component-count agreement against the full-sample fit (Table 50).

Table 50: Covariance-stability audit under 90% subsampling (100 refits/class). **Ratio** is $\min_c N_c/d$; **corr. dev.** is mean absolute pairwise-correlation deviation; **Worst** is the largest per-class correlation deviation; **eig. aff.** is top-3 eigenspace affinity (1.0 = identical subspace); k**-stab.** is BIC component-count agreement.

Dataset	Ratio	Corr. dev.	Worst	Eig. aff.	k-stab.
Iris	12.50	0.037	0.049	0.998	0.65
Wine	3.69	0.036	0.041	0.976	0.90
BreastCancer	7.07	0.014	0.015	0.999	0.78
Optdigits	1.45	0.068	0.075	0.960	1.00
Diabetes	33.50	0.020	0.023	0.929	0.50
Vehicle	11.06	0.015	0.018	0.979	0.84
Segment	18.33	0.025	0.029	0.977	0.45
Banknote	152.50	0.008	0.008	1.000	0.95
Climate-Model	2.30	0.012	0.016	0.652	1.00
Ionosphere	3.71	0.046	0.058	0.930	1.00
Seeds	10.00	0.023	0.025	0.999	0.87
Dermatology	0.59	0.129	0.213	0.749	1.00
Pendigits	5.31	0.041	0.051	0.929	0.44
Waveform	8.10	0.014	0.015	0.907	1.00

Mean correlation deviation is 0.035 across the 14 datasets. Datasets with $\min_c N_c/d \geq 4$ average 0.022, those below average 0.058. The single worst class (0.213) is Dermatology, where the ratio is 0.59, exactly the small-N_c regime already flagged in App. M. The dependence structure is stable precisely where SDS claims applicability, and N_c/d is a usable a priori stability diagnostic.

L.5 Generative-Model-Specific Membership Attacks

Table 9 and Table 18 audit SDS against nearest-neighbor and query-adaptive membership attacks. Here we add a density-ratio attack family built specifically for synthetic-data releases: DOMIAS (van Breugel et al., 2023) (KDE density ratio against a disjoint population reference), a LOGAN-style density-only variant, and a discriminative density-ratio attack (gradient boosting), 3 seeds/dataset, evaluated in an 8-component PCA space fit on the reference sample (Table 51).

SDS stays near the same chance level as the NN-MIA audit reports (worst dataset 0.597), while the identical attacks reach mean AUC 0.681 against ForestDiffusion on the 5-dataset contrast: the attack family clearly can detect a memorizing generator. It simply finds nothing to detect in SDS releases.

L.6 Generalization to a Second In-Context Model (TabICL)

§6 evaluates the TabPFN-2.5 weights through our re-implementation and through the official package. We add TabICL (Qu et al., 2025), an independently developed in-context tabular model with a

Table 51: DOMIAS / LOGAN-density / classifier density-ratio attack AUC on SDS, 14-tier. FDiff column (5-dataset contrast) uses the DOMIAS variant.

Dataset	DOMIAS	LOGAN-D	Clf-ratio	FDiff (DOMIAS)
Iris	0.534	0.579	0.536	0.666
Wine	0.552	0.578	0.544	0.706
BreastCancer	0.531	0.537	0.515	0.666
Diabetes	0.535	0.532	0.524	0.717
Vehicle	0.524	0.515	0.514	–
Segment	0.519	0.509	0.504	–
Banknote	0.531	0.539	0.524	–
Climate-Model	0.517	0.502	0.532	–
Ionosphere	0.507	0.487	0.519	–
Seeds	0.570	0.546	0.550	0.652
Dermatology	0.597	0.556	0.592	–
Pendigits	0.584	0.539	0.553	–
Waveform	0.501	0.478	0.522	–
Optdigits	0.551	0.549	0.529	–
Mean	0.539	0.532	0.533	0.681

different architecture (column-then-row attention) and pretraining, on the full 14-tier (3-fold $\times$ 2 seeds), using the identical SDS releases that a TabPFN client would send (Table 52). App. L.7 extends this study to the 2026 generation of foundation models.

Table 52: TabICL, real vs. SDS context, 14-tier (3-fold $\times$ 2 seeds).

Dataset	Real	SDS	Δ
Iris	0.970	0.973	$+0.3\%$
Wine	0.997	0.989	-0.8%
BreastCancer	0.975	0.966	-1.0%
Diabetes	0.766	0.754	-1.2%
Vehicle	0.878	0.801	-7.7%
Segment	0.986	0.936	-5.0%
Banknote	1.000	0.999	-0.1%
Climate-Model	0.904	0.908	$+0.5\%$
Ionosphere	0.956	0.927	-2.8%
Seeds	0.962	0.952	-1.0%
Dermatology	0.980	0.973	-0.7%
Pendigits	0.988	0.972	-1.6%
Waveform	0.871	0.845	-2.6%
Optdigits	0.983	0.966	-1.7%
Mean	0.944	0.926	$\mathbf{-1.8}\%$

13 of 14 datasets stay within 5%. The per-dataset pattern mirrors TabPFN's own (Vehicle is the worst case under both models), which supports transferring one release across readers, though it does not show that the gap is independent of the reader. The only setting tuned to TabPFN, the oversampling factor α, transfers unchanged.

L.7 Generalization to the 2026 Generation of Tabular Foundation Models

We repeat the protocol of App. L.6 at the full headline protocol (5-fold $\times$ 3 seeds, α=2, identical SDS releases) on the current generation of in-context tabular foundation models: TabPFN-2.5 (Grinsztajn et al., 2025), TabICLv2 (Qu et al., 2026), LimiX-16M (LimiX Team, 2025), and Google's TabFM (Google Research, 2026), plus one LLM-class tabular predictor, TabuLa-8B (Gardner et al., 2024). Table 3 (in §6) reports mean accuracy with the real context versus the SDS synthetic context.

The four full-context models reproduce the pattern seen for the headline TabPFN reader and TabICL v1 (App. L.6): gaps Δ of -1.1% to -1.9%, twelve to thirteen of fourteen datasets within 5%, and the same single worst dataset, Vehicle (for the TabPFN reader, -5.0% at α=8 and -4.3% with a

tuned GMM-only generator, App. L.3). Across these six readers of five independently pretrained models, both the mean gap and the worst-case dataset barely move, which supports transferring one release across readers, although it does not show that the gap is independent of the reader. TabuLa-8B, evaluated at its 32-shot protocol, extends the pattern to an LLM-class reader at much lower absolute accuracy: the gap stays moderate (-2.5% mean, $11/14$ within 5%), so SDS contexts remain usable even for a model that consumes rows as text.

L.8 A Numerically Informative Pinsker Bound in the Capped-Release Regime

The bound's numerical behavior splits into two regimes, both governed by Theorem 2: the *oversampled* regime used throughout §6, where the release size scales with the dataset, and the *capped-release* regime SDS already uses for large datasets (Table 33 caps the context at 500 records regardless of N). Since a balanced adversary's advantage cannot exceed 0.5, a ceiling above 0.5 carries no information, so in the first regime the audited NN-MIA AUC of 0.49–0.60, not the bound, underwrites the privacy claim.

§5 shows the Pinsker bound of Corollary 2 is vacuous (> 0.5) on 13 of the 14 datasets in Table 5, because the released context $m{=}\alpha N$ scales with the dataset: cumulative KL $= m\times$(per-record KL) stays roughly constant as N grows rather than shrinking. Recomputing the same, already-audited per-record KL estimates from Table 5 at a fixed $m{=}200$ (independent of N) instead gives Table 53:

Table 53: Pinsker-style diagnostic recomputed at a fixed released-context size $m{=}200$, using the per-record KL already estimated for Table 5. Four of 14 datasets sit below the vacuousness threshold (0.5) with no new estimation. These are $\widehat{\Gamma}$ plugged into the theorem's formula, not the supremum the theorem requires.

Dataset	**Diag. at $m{=}\alpha N$**	**Diag. at $m{=}200$**	**Non-vacuous?**
Iris	0.599	0.489	Yes (marginal)
BreastCancer	0.748	0.314	Yes
Seeds	0.482	0.333	Yes
Waveform	1.022	**0.145**	**Yes**

Waveform is the clean case: its observed NN-MIA AUC is 0.55 (Table 9), advantage ≈ 0.05. The recomputed diagnostic of 0.145 is loose ($\approx 3\times$) but numerically non-vacuous, using zero new estimation. It rests on the same sampled $\widehat{\Gamma}$ as Table 5 rather than a supremum over neighbours, so it stays a diagnostic and not a certified ceiling (Table 5).

To confirm this mechanism is real rather than a coincidence of these four datasets, we ran a controlled sweep: one fixed synthetic distribution ($d{=}10$, $C{=}2$), fixed $m{=}200$, sweeping the per-class fitting size N_c from 20 to 20,000, using the identical replace-one KL protocol as Table 5 (Table 54).

Table 54: Controlled sweep: fixed $m{=}200$, swept N_c. **Emp. adv.** is the empirically measured NN-MIA advantage at that N_c. The diagnostic exceeds the measured attack at every point tested, which is consistency rather than a guarantee. Two further small-N_c points with vacuous diagnostics (10.48 and 9.69) are omitted from the table.

N_c	**Diagnostic**	**Non-vacuous?**	**Emp. adv.**
20	3.471	No	0.091
100	0.581	No	0.048
500	1.428	No	0.016
1,000	0.161	**Yes**	0.005
2,000	0.108	Yes	0.005
5,000	0.076	Yes	0.002
10,000	0.037	Yes	0.008
20,000	0.028	Yes	0.001

The diagnostic exceeds the measured attack at all 10 sweep points tested (including the noisy small-N_c points not shown, where a single replace-one perturbation can flip a borderline BIC component choice, the same phenomenon noted for Table 5), crosses into non-vacuous territory around $N_c \approx 1{,}000$, and

tightens like $O(\sqrt{m}/N_c)$ thereafter. Both regimes are governed by the same theorem (Theorem 2). The small-N, oversampled regime used for the headline results stays structural, exactly as §5 states, while the capped-release regime SDS already uses for large datasets (Table 33) yields a computable, non-vacuous diagnostic.

L.9 Per-Stage Timing

SDS's own fit-and-sample step is CPU-only and runs in under one second per fold across the 14-dataset tier (per-class BIC-GMM and Gaussian-copula fits, $N \leq 11{,}000$, $d \leq 64$). The GPU is used only for TabPFN inference and the ForestDiffusion baseline. ForestDiffusion training takes $\approx$4,000 s/fold at d=57 and is infeasible at $d \geq 216$ within our compute budget, which is why its coverage is narrower than SDS's in Table 19. Candidate self-prediction selection (§4, K=10) adds one local TabPFN forward pass per candidate, dominated by TabPFN inference cost rather than SDS's own fit time. App. J.4 shows this step is itself optional (Table 40).

L.10 Geometry Controls

The geometry metric of §6 reconstructs the released cloud by MDS and aligns it to the real records by index. Running the same protocol on a row-permuted copy of the real records, a release that exposes every training record verbatim, gives what Table 55 reports. It matches the SDS column, so the metric measures alignment rather than exposure and we do not read it as a privacy result. The correspondence-free evidence for this axis is the near-record rate, which on the headline tier leaves synthetic records nearly as far from training members as from held-out non-members.

Table 55: The synthetic-side geometry protocol applied to the core six (3 seeds), SDS against a row-permuted copy of the real records.

Dataset	SDS	Real rows, permuted
Iris	0.020	0.027
Wine	0.048	0.053
BreastCancer	0.059	0.052
Multi3_Med	0.064	0.059
Multi5_Med	0.115	0.121
HighDim_50	0.229	0.235
Mean	0.089	0.091

L.11 Simple Generators and Noise-Perturbed Releases

Table 56 prices the two obvious alternatives to SDS on the 14-dataset numeric tier: dropping a component of the generator, and perturbing the real rows instead of synthesising new ones.

L.12 Candidate-Selection Sweep

Self-prediction selection draws K candidate contexts and keeps the one that best re-predicts the owner's real labels. Table 57 sweeps K on the 14-dataset numeric tier (2 seeds, 3 folds). Utility is flat in K, and the per-dataset difference between K=10 and K=1 ranges from -0.95 to $+0.78$ points with no consistent sign, so the gain is indistinguishable from fold noise while the local cost grows linearly in K. Membership under the two settings is audited separately at the headline protocol in Table 40, which is the comparison we rely on.

L.13 Differential-Privacy Budget Control

The DP baselines of §6 lose most of their accuracy, which invites the question of whether the implementation is at fault rather than the privacy noise. Table 58 sweeps the budget of the class-conditional Gaussian mechanism on the 14-dataset numeric tier. At an effectively unbounded budget the mechanism recovers to within 2.6 points of the real context, so it can fit these datasets, and the loss at ε=1 is attributable to the noise rather than to a configuration error.

Table 56: Simple-generator and noise-perturbation baselines on the 14-dataset numeric tier (3 seeds, utility 3-fold, member set capped at 400 rows). Absolute values are not comparable with the headline audit of Table 9, which uses 5-fold × 3 seeds on the uncapped member set. The comparison here is between releases within the table. Lower is better on every column except Δ_{util}, where closer to zero is better. Bold marks the best utility. Perturbing real rows never reaches SDS-level membership protection: the closest setting, σ=1, still leaks more (NN-MIA 0.602 against 0.562) while costing 3.7× the accuracy. The SDS row and the σ=0.25 row coincide at three decimals on attribute R^2 (0.6037 against 0.6042) and on the near-record rate (0.1925 against 0.1933).

Release	Δ_{util} (pp)	**NN-MIA**	**Attr R^2**	**Near-rec.**
SDS ensemble (default)	$\mathbf{-1.38}$	0.562	0.604	0.193
BIC-GMM only	−2.09	0.556	0.622	0.176
Copula only	−1.85	0.562	0.571	0.187
One Gaussian per class	−2.61	0.550	0.623	0.167
Real rows + $\mathcal{N}(0, 0.25^2)$	−1.60	0.839	0.604	0.193
Real rows + $\mathcal{N}(0, 0.50^2)$	−2.33	0.726	0.575	0.119
Real rows + $\mathcal{N}(0, 1.00^2)$	−5.09	0.602	0.494	0.060
Real rows resampled	−1.66	0.938	0.616	0.900

Table 57: Candidate selection on the 14-dataset numeric tier. Selection leaves utility unchanged while the local cost grows linearly in K. The headline runs use K=1 (no selection), since the accuracy difference is within fold noise and the leakage audit of Table 40 is flat.

Candidates	Δ_{util} (pp)	**local TabPFN calls**
K=1 **(default)**	−1.25	1
K=5	−1.27	5
K=10	−1.27	10

L.14 Privacy–Utility Frontier Against a DP Mechanism

The DP baselines of §6 are reported at fixed budgets, which shows that they cost accuracy but not whether a better-tuned budget would dominate SDS. Table 59 answers that directly. We sweep ε on the class-conditional Gaussian mechanism over the 14-dataset numeric tier (3 seeds, member set capped at 400 rows) and measure utility and all three leakage axes on the *same* split, so the two axes are directly comparable. The SDS row uses the identical protocol. *Excess* is the attribute score minus that of the same attacker trained on an independent population sample of equal size, which is the quantity a deployment actually cares about.

The mechanism never dominates SDS. Its membership advantage sits at or below SDS's at every budget, but so does its accuracy, by 1.3 points at ε=100 and by more than 44 points at ε=0.5. To come within 1.3 points of SDS's accuracy it needs ε=100, at which the budget is no longer meaningfully private, its membership advantage is only 0.011 below SDS's, and its attribute excess is *higher* than SDS's. At no budget is the mechanism better than SDS on both axes at once.

L.15 Repeated Releases Beyond Twenty

Table 60 extends the composition audit to 200 pooled releases on all fourteen headline datasets. Growth is modest almost everywhere, at most +0.075 AUC over the full range on twelve datasets. The two exceptions, shown in bold, are Iris and Seeds, the only datasets that are both low-dimensional and small. They are the same cases the distance-to-closest-record rule already flags, so the practical guidance is a dimension-and-size condition rather than a universal release count.

L.16 Gen-LRA Against the Baseline Releases

Table 61 runs the no-box Gen-LRA attack against every release of App. L.11 except the σ=0.5 perturbation. The reimplementation reproduces the audit of App. F.19 to within 0.01 on SDS and 0.02 on the memorizing control, so the two agree. Its value here is the extension: a lightly noise-perturbed

Table 58: Accuracy cost of the class-conditional Gaussian DP mechanism as a function of the privacy budget, 14-dataset numeric tier (2 seeds, 3 folds).

ε	Δ_{util} (pp)
1	-37.89
10	-10.05
100	-2.80
10^6 (effectively non-private)	-2.60

Table 59: Privacy–utility frontier on the 14-dataset numeric tier (3 seeds, member set capped at 400, α=2). Δ_{util} is synthetic-context minus real-context TabPFN accuracy on the same holdout. NN-MIA is the attack of Eq. (2) (chance 0.5). *Attr* is Ridge attribute inference, *pop* the same attacker trained on an independent population sample, and *excess* their difference. *Near-rec* is the fraction of members with a synthetic neighbour within 0.5 standardised units. Lower is better on every column except Δ_{util}, where closer to zero is better.

Release	Δ_{util} (pp)	**NN-MIA**	**Attr**	**pop**	**excess**	**Near-rec.**
SDS ensemble	$\mathbf{-1.40}$	0.562	0.604	0.520	$\mathbf{+0.084}$	0.193
DP ε=0.5	-45.62	0.508	-0.220	0.520	-0.740	0.044
DP ε=1	-40.32	0.517	0.159	0.520	-0.361	0.063
DP ε=2	-31.74	0.534	0.432	0.520	-0.089	0.073
DP ε=5	-21.55	0.546	0.573	0.520	$+0.053$	0.129
DP ε=10	-13.53	0.548	0.607	0.520	$+0.087$	0.155
DP ε=20	-7.80	0.550	0.618	0.520	$+0.098$	0.161
DP ε=50	-3.26	0.551	0.622	0.520	$+0.102$	0.162
DP ε=100	-2.65	0.551	0.622	0.520	$+0.102$	0.163

release of the real rows (σ=0.25) is as detectable as a deliberately memorizing generator, while every synthetic generator stays near chance. A second attack family therefore reaches the same conclusion as the nearest-neighbour audit.

L.17 Attribute Inference Against a Population Control

Attribute inference is best read as the *excess* over what an adversary could obtain without the release at all. To measure it we refit the attribute attacker on an independent same-size sample of the same population and evaluate it on the same real targets. The excess is small, and it shrinks as the attacker strengthens: $+0.087$ with Ridge, $+0.050$ with gradient boosting, $+0.025$ with a random forest and $+0.021$ with Extra Trees, the strongest attacker we run. The frontier of App. L.14 reproduces this independently at $+0.084$ for Ridge on the same tier. Most of what the raw R^2_{attr} measures is therefore population structure the adversary already has, not exposure of the released rows.

Stronger attackers. Ridge is linear, so it may understate the leakage available from a generator that preserves nonlinear structure. Table 62 repeats the population comparison with three nonlinear attackers. The gap between an SDS-trained and a population-trained attacker *shrinks* as the attacker strengthens, from 0.087 with Ridge to 0.021 with Extra Trees, because the stronger attacker also extracts more from the population sample. The largest per-dataset Ridge gap, $+0.497$ on Ionosphere, falls to $+0.053$ under Extra Trees, so it reflects a degenerate linear fit on the population sample rather than exposure of the training set.

M Additional Limitations

When not to use SDS as-is. Explicit decision rules from the paper's diagnostics: (a) $d \lesssim 10$, apply the DCR filter (App. F.25) or treat near-record leakage as a live risk, and for pooled releases use the band-pass rule of App. J.6. (b) $N_c/d < 4$ for any class, covariance estimation is measurably less stable (App. L.4) and Ledoit–Wolf shrinkage compensates only partially. (c) population-level attribute inference on regulated columns is unacceptable for the deployment, use the hardening of

Table 60: NN-MIA AUC against the number of pooled successive releases (2 seeds). N_{mem} is the member-set size. Bold rows are simultaneously low-dimensional and small.

Dataset	N_{mem}	d	k=1	k=5	k=20	k=50	k=100	k=200
Banknote	686	4	0.531	0.524	0.519	0.539	0.536	0.544
BreastCancer	284	30	0.499	0.502	0.508	0.518	0.518	0.519
Climate-Model	270	20	0.521	0.570	0.559	0.553	0.574	0.596
Dermatology	179	34	0.600	0.613	0.619	0.624	0.630	0.640
Diabetes	384	8	0.524	0.540	0.542	0.552	0.553	0.557
Ionosphere	175	34	0.516	0.521	0.522	0.523	0.523	0.524
Iris	75	4	0.539	0.613	0.690	0.692	0.704	**0.749**
Optdigits	500	64	0.586	0.584	0.599	0.608	0.615	0.616
Pendigits	500	16	0.553	0.566	0.576	0.585	0.588	0.595
Seeds	105	7	0.592	0.640	0.677	0.740	0.770	**0.764**
Segment	1000	19	0.527	0.530	0.535	0.534	0.540	0.540
Vehicle	423	18	0.541	0.535	0.552	0.555	0.567	0.572
Waveform	500	40	0.576	0.582	0.602	0.611	0.614	0.613
Wine	89	13	0.616	0.633	0.645	0.666	0.660	0.660
Mean			0.552	0.568	0.582	0.593	0.599	0.606

Table 61: Gen-LRA on the 14-dataset numeric tier (3 seeds, KDE in at most eight principal components). Chance is 0.5. Perturbing real rows leaks as much as the memorizing positive control.

Release	**Gen-LRA AUC**	**advantage**
SDS ensemble	0.552	0.052
BIC-GMM only	0.546	0.046
Copula only	0.552	0.052
One Gaussian per class	0.538	0.038
Real rows $+\,\mathcal{N}(0, 0.25^2)$	0.709	0.209
Real rows $+\,\mathcal{N}(0, 1.00^2)$	0.558	0.058
Real rows resampled	0.751	0.251
Memorizing generator (control)	0.706	0.206

§6 or suppress those columns. Hardening alone does not eliminate this axis. (d) many releases from a single small dataset over time, the tested attack increases modestly on most datasets out to 200 releases but substantially on Iris and Seeds (App. L.15). (e) selection needs local candidate scoring, so with no local compute use the RF-proxy selector (App. J.4) or keep the default K=1, which costs no measurable accuracy (Table 57). (f) the labels are an exact deterministic rule over the features (parity or XOR-type interactions, board-game legality, hard threshold rules on software or engineering metrics), where the smooth Gaussian-mixture/copula generator cannot reproduce the sharp decision surface and a rule-learner or the real context should be used instead.

1. **Honest-but-curious model.** Stronger threat models (adaptive test queries, model modification) are not covered.
2. **Label distribution.** Class fractions N_c/N reveal label prevalence to the server, mitigated by the Laplace mechanism in §6.
3. **Small N_c.** When $N_c < d$, covariance estimation degenerates. Ledoit–Wolf shrinkage compensates only partially. App. L.4 quantifies this directly: mean correlation deviation under subsampling is 0.022 when $\min_c N_c/d \geq 4$ vs. 0.058 below that ratio, so N_c/d is a usable a priori stability check.
4. **Benchmark scope.** The mixed-type evaluation covers 15 CC18/OpenML datasets (Table 13) and 31 of 36 TabZilla hard-benchmark datasets (McElfresh et al., 2023) (5 excluded: audiology and lymph for a class with fewer than six members, colic for $N < 100$, albert and guillermo for OpenML load failures). The numeric evaluation uses a custom 14-dataset suite rather than a standard benchmark like TabArena.
5. **Deep-baseline coverage.** Among deep baselines, TVAE is tested on all 14 repeated-tier datasets (SDS wins 13/13, Wilcoxon $p = 0.00012$, Table 15). TabDDPM (Kotelnikov et al., 2023) has no pip-installable package and uses an MLP with $O(10^5$–$10^6)$ parameters

Table 62: Attribute inference against a population control, across attacker classes, 14-dataset numeric tier (3 seeds). Per-column R^2 clipped at -1 before averaging.

Attacker	Trained on SDS	Trained on population	Δ
Ridge (headline attacker)	0.604	0.517	+0.087
Random Forest	0.608	0.583	+0.025
Extra Trees	0.613	0.593	+0.021
HistGradientBoosting	0.578	0.527	+0.050

requiring $N \geq$ 3,000–5,000 to converge (Kotelnikov et al., 2023, Table 1), above our $N_c \in [30, 350]$ regime where even TVAE incurs a -14.1% gap. ForestDiffusion (Jolicoeur-Martineau et al., 2024) keeps pace on accuracy (it wins $8/14$ per-dataset comparisons against SDS, Table 7) but its NN-MIA AUC is 0.73 (Table 9), consistent with the memorisation risk of diffusion-based generative models (Somepalli et al., 2023).

6. **Near-record leakage.** For Iris (d=4), 58.8% of records have a synthetic neighbor within ε=0.25. This is concentrated in low dimensions (Table 32) but not confined to them: kc1 (d=21) reaches a near-record fraction of 0.51 at α=2 (Table 47). The DCR filter is the designated mitigation: on the flagged Iris case it halves NN-MIA ($0.117 \rightarrow 0.053$) at no accuracy cost (Table 32). Concrete rule: apply the filter when $d \lesssim 10$, or when the near-record fraction at ε=0.5 measured on the training set crosses a chosen threshold.
7. **Correspondence-free geometry attacks.** The geometry claim covers the evaluated MDS-plus-Procrustes family (App. F.4). Reconstruction attacks that jointly optimize the record correspondence, such as Gromov–Wasserstein alignment, are not evaluated and remain an open extension.
8. **Repeated and heterogeneous releases.** Pooling many releases of the same small dataset compounds membership leakage on the smallest datasets (App. L.2). Out to 200 releases the effect stays at most +0.075 AUC on twelve of the fourteen datasets but carries Iris to 0.749 and Seeds to 0.764, the two that are both low-dimensional and small (App. L.15). On the TabZilla-hard tier the nearest-neighbour attack reaches 0.688 on one-hundred-plants-texture, the highest single-release value we measure, while the tier mean stays at 0.522 (App. I).
9. **Deterministic-rule labels.** SDS's generator assumes each class-conditional density is well approximated by a Gaussian mixture and copula, i.e. smooth statistical structure. When the label is instead an exact deterministic function of the features (e.g. parity/XOR interactions, board-game rules, or hard metric thresholds), the sampled context smooths away the precise interactions the decision boundary depends on and utility degrades sharply.
10. **Continuous-only theory.** Mixed-type tables violate the absolute-continuity assumption. The discrete pipeline (Chow-Liu with per-feature flip noise and a GBM label refiner) handles categorical columns natively, but the non-reproduction property of a continuous generator ($\mathbb{P}[x_{\text{synth}} = x_{\text{real}}] = 0$ under an absolutely continuous p_G) no longer holds for discrete features. We audit MIA AUC empirically instead and verify it stays at chance (Table 17).
11. **Stronger audits and modern DP baselines.** Our membership audit spans nearest-neighbor, query-adaptive, shadow-model, density-ratio (DOMIAS (van Breugel et al., 2023)), and the no-box local-likelihood-ratio attack Gen-LRA (Ward et al., 2025) (App. F.19, L.16). Joint-structure reconstruction that couples per-feature posteriors on a dependency graph (Golob et al., 2026), read out with rarity-weighted and memorization-versus-population diagnostics, is stronger and is the natural next audit, alongside correspondence-free geometry attacks. On the baseline side, benchmarking against a modern DP latent-diffusion synthesizer (DP-TLDM (Zhu et al., 2024)) would place SDS's non-private, utility-faithful release on the current privacy–utility frontier.