ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

Generative AI for Credit Risk Management: Risk-Conditional Diffusion with Holdout-Calibrated Release Gates for Empirically Privacy-Audited Synthetic Credit Data

Nathan Turner, Hannah Tan, Anika Qian *

School of Information and Computer Sciences, University of California, Irvine, CA, United States

* Corresponding Author: anikaqian09@uci.edu

Abstract

Generative AI is increasingly used to share and augment confidential credit data, but high-fidelity generators can reveal which borrowers they were trained on. We study this tension on three public credit risk datasets and propose PG-Diff (proximity-gated diffusion), which couples a risk-conditional tabular diffusion model, combining mass-point encoding of zero-inflated amounts with classifier-free guidance on the default label, with a generator-agnostic release gate. The gate enforces absolute proximity quotas and a relative proximity balance, both calibrated on held-out real records so that records drawn independently of the training data pass them in expectation. Adopting the fidelity, utility and privacy evaluation framework of Yang and Yang, extended with low-false-positive membership inference and a model-risk agreement metric, we compare SMOTE, Gaussian copula, TVAE, CTGAN and diffusion releases over up to three random splits. Conditional diffusion reaches train-on-synthetic AUCs of 0.751, 0.845 and 0.748 on Taiwan, Give Me Some Credit and German (real-data models: 0.762, 0.834, 0.778), gives the most realistic records among the deep generators, and shows membership-inference AUCs close to 0.5 on the two larger datasets. Interpolation-based SMOTE attains similar utility but leaks heavily (attack AUC up to 0.76, true-positive rate at 1% false positives up to 0.58); the gate lowers that rate by 43% on average and flags SMOTE through relaxation rates of up to 97%. For non-interpolating generators the gate costs at most 0.039 AUC, and in a memorization stress test it moves the DCR share from 0.595 to 0.472. Holdout-calibrated release gates thus provide a generator-agnostic empirical safeguard that complements formal mechanisms such as differential privacy.

Keywords

Generative AI, credit risk, synthetic data, diffusion models, membership inference, empirical privacy auditing, model risk management.

1. Introduction

Financial institutions are turning to generative artificial intelligence (AI) to deal with a structural problem in risk management: the data that matter most for credit, fraud and market risk models are also the data that are hardest to move. Synthetic records produced by deep generative models can be shared across business lines, handed to vendors and research partners, and used to rebalance rare default or fraud events [1]. Generative sequence models simulate market scenarios with realistic stylized facts [2], and large language models are being adapted to financial text for analysis and reporting [3]. In a recent review, Yang and Yang [4] surveyed GAN-, VAE- and diffusion-based synthesis for financial analytics and concluded that no single paradigm dominates across use cases and that evaluation frameworks combining statistical fidelity with task-specific utility remain an open research priority.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

For credit risk, the value of synthetic data hinges on two properties that pull in opposite directions. A synthetic portfolio must be faithful enough for probability-of-default (PD) models trained on it to rank real borrowers correctly, and it must not reveal whether a particular borrower was in the training data. The second property cannot be taken for granted. Synthetic data are not anonymous by construction [5]; diffusion models, currently among the strongest tabular generators [6, 7], memorize and regenerate training examples [8], and the same behavior has recently been documented for tabular diffusion [9]. Supervisory guidance on model risk management further requires that models developed with such data be validated and their behavior understood [10], which raises the bar on what a synthetic release must preserve.

Two families of safeguards exist. Differentially private (DP) training [11–13] provides formal guarantees but often costs substantial utility on mixed-type, imbalanced financial tables [14]. Post-hoc similarity filters, which discard synthetic records that look too close to real ones, are popular in practice, yet their thresholds are usually ad hoc and they are often judged by the same similarity statistics they optimize; Ganev and De Cristofaro [15] show that this practice can be badly misleading. What is missing is a release mechanism whose thresholds are calibrated against an explicit reference, namely the behavior of genuinely fresh records, and whose effect is audited with attacks it was not tuned on.

This paper addresses that gap for credit risk data. We pair a risk-conditional tabular diffusion model with a generator-agnostic release gate calibrated on a held-out real sample. The gate follows a principle borrowed from holdout-based assessment of synthetic data [16]: a released record should be no closer to the training data than an independent real record would be. The main contributions are as follows.

A risk-conditional diffusion generator for mixed-type credit tables that combines mass-point encoding of zero-inflated monetary attributes with classifier-free guidance [17] on the default label, so that releases have an exact default rate and controllable class separation.

A holdout-calibrated release gate with two constraints, absolute proximity quotas (HPQ) and relative proximity balance (HBP), together with an exchangeability argument showing that records drawn independently of the training set satisfy both constraints in expectation.

A three-axis evaluation of fidelity, utility and privacy, extended with low-false-positive membership inference [18, 19] and a model-risk agreement metric, applied to three public credit datasets and five generator families over three random splits.

Empirical evidence on where generative AI helps and where it leaks in credit risk work, including the use of the gate as an admissibility audit that flags generators whose outputs are structurally too close to their training data.

Section II reviews related work, Section III presents the method, Section IV describes the experimental protocol, Section V reports the results, and Section VI discusses applications, open challenges and future directions.

2. Related Work

2.1. Generative Models for Financial Data

Deep generative models for tabular data evolved from adversarial and variational synthesizers to diffusion. CTGAN and TVAE [20] introduced mode-specific normalization and training-by-sampling to cope with multimodal continuous columns and imbalanced categories, while copula-based synthesizers such as those in the Synthetic Data Vault [21] model dependence through a latent Gaussian. Denoising diffusion probabilistic models [22] were adapted to mixed-type tables by TabDDPM [6], which combines Gaussian diffusion on quantile-transformed numerical features with multinomial diffusion on categorical ones, and by TabSyn

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

[7], which runs score-based diffusion in the latent space of a variational autoencoder. In finance, generative models have been used to synthesize return series [2], to rebalance default and fraud data [1], and to build domain-adapted language models [3]. Symmetry-aware credit risk modeling demonstrates the relevance of learning balanced representations that account for invariance in heterogeneous borrower information [31].

Yang and Yang [4] organize this literature into adversarial, latent-variable and score-based lineages. Two of their observations shape our design: diffusion models offer better-calibrated tails, which matters when extreme quantiles drive capital calculations, and they integrate naturally with classifier-free guidance for conditioning on risk categories without architectural changes. Our generator builds on both points. Zuo et al. [14] benchmarked copula, diffusion, CTGAN and TVAE synthesizers on financial tables, including two of the datasets used here, and found that shadow-model membership attacks rarely exceeded chance. We complement their study with record-level attacks that are known to be stronger against synthetic data [19] and with a mitigation mechanism.

2.2. Privacy Risks of Synthetic Data

Membership inference asks whether a target record was part of a model's training set [18]. For synthetic data, Stadler et al. [5] showed that releasing generated records instead of real ones does not by itself prevent linkage and inference, and DOMIAS [19] showed that contrasting the synthetic density with a reference density exposes local overfitting that earlier attacks missed. Carlini et al. [8] extracted training images from diffusion models, and Fang et al. [9] found that memorization in tabular diffusion grows with the number of training epochs, which they mitigate through augmentation during training. DP training [12, 13] bounds such leakage formally [11] but degrades utility on small, imbalanced tables. Ganev and De Cristofaro [15] warn that similarity-based metrics and filters can be satisfied while records remain recoverable. We take this criticism seriously: our gate is calibrated against fresh data rather than fixed thresholds, it is evaluated with attacks it does not optimize, and we position it as a complement to DP rather than a substitute.

2.3. Evaluating Synthetic Data

Train-on-synthetic, test-on-real (TSTR) utility, statistical fidelity and privacy auditing are the standard evaluation axes for tabular synthesizers [6, 7]. Fidelity is commonly measured with marginal and pairwise statistics, classifier two-sample tests and precision/recall in feature space [23]. Privacy is often summarized by the distance to the closest record (DCR); TabSyn [7] reports the DCR share, i.e., the probability that a synthetic record is closer to the training set than to an equally sized holdout, and Platzer and Reutterer [16] argue more generally for holdout-based reference distributions. We reuse these quantities but keep the holdout used for calibrating the gate strictly separate from the holdout used for evaluation.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

3. Methodology

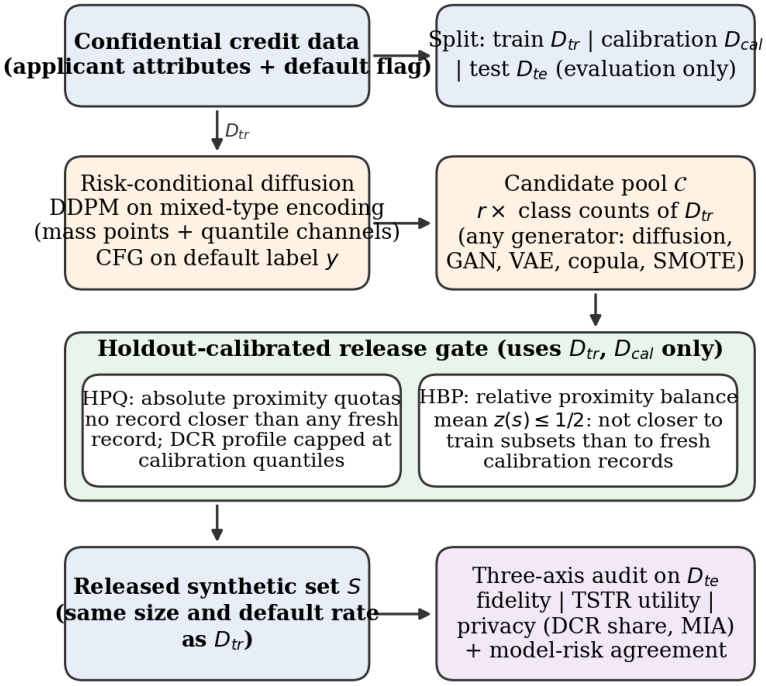


Figure 1. Overview of PG-Diff. A risk-conditional diffusion model produces a candidate pool; a release gate calibrated on held-out real records selects the published set; all claims are audited on a separate test split.

3.1. Setting and Threat Model

Let $D = \{(x_i, y_i)\}$ be a confidential credit portfolio, where x collects numerical and categorical applicant or account attributes and $y \in \{0, 1\}$ indicates default. The data holder partitions D into a training split D_{tr}, a calibration split D_{cal} and a test split D_{te}. Generators see only D_{tr}, the release gate sees D_{tr} and D_{cal}, and D_{te} is reserved for evaluation (Fig. 1). The goal is to publish a synthetic table S with $|S| = |D_{tr}|$ and the same number n_c of records in each class c, so that PD models can be developed and validated on it. The adversary observes S, knows the schema, and holds a reference sample from the same population, which is the standard assumption of DOMIAS [19]; it tries to decide whether a target record belongs to D_{tr}. We do not give the adversary access to the generator itself.

3.2. Mixed-Type Encoding with Mass Points

Credit tables contain monetary attributes with large point masses; in the Taiwan data every bill and repayment attribute has an atom at zero. A quantile transform maps such an atom to a single value that a Gaussian diffusion model reproduces poorly: in a pilot run with plain quantile encoding, the zero-inflated repayment columns had Kolmogorov–Smirnov distances of 0.17–0.27 to the held-out data, the largest marginal errors in the table. We therefore encode every numerical attribute j with (i) a one-hot block over its mass points, defined as the at most three values that each carry at least 5% of the training records, plus a "continuous" slot, and (ii) one channel holding the Gaussian quantile transform of the remaining values (zero for mass-point records). Categorical attributes are one-hot encoded with entries in {−1, +1}. Decoding takes the arg max of each block, inverts the quantile transform, clips to the observed range and rounds integer attributes. The encoding is in the spirit of CTGAN's mode-specific normalization [20], but it keeps a single Gaussian diffusion process over the whole vector instead of TabDDPM's separate multinomial process [6].

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

3.3. Risk-Conditional Diffusion

Let x_0 denote an encoded record. The forward process corrupts it as

$$x_t = \bar{\alpha}_t^{1/2} x_0 + (1 - \bar{\alpha}_t)^{1/2} \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I) \tag{1}$$

With a cosine schedule for $\bar{\alpha}_t$ over $T = 1000$ steps. A multilayer perceptron ε_θ receives x_t, a sinusoidal embedding of t and a learned embedding of the label, and is trained with the denoising objective of [22]

$$\mathcal{L}(\theta) = \mathbb{E} \, \|\varepsilon - \varepsilon_\theta(x_t, t, \tilde{y})\|^2 \tag{2}$$

Where $\tilde{y}$ equals y with probability $1 - p_u$ and a null token $\emptyset$ otherwise ($p_u = 0.1$). The null token trains an unconditional model in the same network, which enables classifier-free guidance [17] at sampling time:

$$\hat{\varepsilon} = (1 + w) \, \varepsilon_\theta(x_t, t, y) - w \, \varepsilon_\theta(x_t, t, \emptyset) \tag{3}$$

Sampling uses 100 strided steps of the generalized update of [24] with $\eta = 1$, which recovers ancestral sampling. Conditioning on y makes it possible to generate exactly n_c candidates per class, which is convenient for releases that must preserve a portfolio's default rate and for targeted minority oversampling. The guidance scale w trades class separation against diversity; $w = 0$ reduces to plain conditional sampling. We fixed $w = 1$ from a pilot on the calibration split of one Taiwan split, before any test-set evaluation, and analyze its effect in Section V.

3.4. Holdout-Calibrated Release Gate

Let φ map a record to a metric space in which numerical attributes are replaced by their empirical CDF values (fitted on D_{tr}) and categorical attributes by one-hot vectors scaled by $1/\sqrt{2}$, so that every attribute contributes at most one unit of squared distance. For a set A, let $d(s, A) = \min_{a \in A} \|\varphi(s) - \varphi(a)\|_2$. The generator first produces a candidate pool $\mathcal{C}$ containing $r \cdot n_c$ candidates of each class; the gate then selects $S \subset \mathcal{C}$. Both of its constraints are calibrated on D_{cal}, whose records are real but were never seen by the generator.

Absolute proximity quotas (HPQ). Let q_k be the ℓk-quantile and d_{min} the minimum of the calibration profile $\{d(h, D_{tr}) : h \in D_{cal}\}$. For levels $\ell_1 < \ldots < \ell_K$, the released records of class c, denoted S_c, must satisfy

$$|\{s \in S_c : d(s, D_{tr}) \le q_k\}| \le \lceil \ell_k \, n_c \rceil \tag{4}$$

For $k = 1, \ldots, K$, and no released record may be closer to D_{tr} than d_{min}. We use levels $\{0.01, 0.02, 0.05, 0.10, 0.20, 0.30\}$. HPQ removes near-copies and caps the lower tail of the DCR distribution at the level observed for fresh data.

Relative proximity balance (HBP). HPQ constrains how close a record is to the training data, but not whether it is closer to the training data than to comparable fresh data. Interpolating generators, for example, can place records at fresh-looking distances that nevertheless lie systematically between training neighbors. Let $m = |D_{cal}|$ and draw B random subsets $T_1, \ldots, T_B \subset D_{tr}$ of size m. For a candidate s define

$$z(s) = (1/B) \, \Sigma_{b=1}^{B} \, \mathbb{1}[d(s, T_b) < d(s, D_{cal})] \tag{5}$$

And require the release to satisfy $(1/|S|)\Sigma_{s \in S} \, z(s) \le 1/2$. We use $B = 8$.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

Proposition 1. If s is drawn from the population distribution independently of D_{tr} and D_{cal}, then $E[z(s)] = 1/2$, and $P(d(s, D_{tr}) \le q_k) \to \ell_k$ as $m \to \infty$.

Proof sketch. D_{cal} and each T_b are independent samples of size m from the same distribution and are exchangeable with respect to an independent s; hence $d(s, T_b)$ and $d(s, D_{cal})$ have the same law, ties are split evenly, and $E[z(s)] = 1/2$. For the second claim, $d(s, D_{tr})$ and $d(h, D_{tr})$ have the same distribution for $h \in D_{cal}$, so the empirical quantiles of the calibration profile converge to those of $d(s, D_{tr})$ by the Glivenko–Cantelli theorem.

Proposition 1 establishes the expected calibration behavior of independently drawn records, rather than guaranteeing that every finite release satisfies both constraints. Because the gate accesses the training data, it acts as an empirical proximity safeguard that complements formal guarantees, as discussed in Section VI.

Selection algorithm. Candidates of each class are visited in random order. A candidate is accepted if it is not closer than d_{min}, if accepting it keeps every quota in (4) satisfied, and if either $z(s) \le 1/2$ or the running mean of z remains at most 1/2. Selection stops at nc records. If the pool is exhausted first, the remaining slots are filled with the admissible-by-HPQ candidates of smallest z, and the number of such relaxed records is reported. A high relaxation rate is itself informative: it means that the generator cannot produce enough records that look fresh, and it is used below as an admissibility audit. The cost is dominated by nearest-neighbor searches and is $O(|\mathcal{C}|(|D_{tr}| + (B + 1)m))$ distance evaluations. As an optional fidelity step (FIR), candidates can instead be visited in an order drawn with probabilities proportional to discriminator-estimated density ratios between D_{cal} and $\mathcal{C}$; we report it as an ablation.

3.5. Evaluation Protocol

We adopt the three evaluation dimensions articulated for financial synthetic data by Yang and Yang [4], namely statistical fidelity, ML efficacy measured by TSTR, and privacy robustness, and we extend each with measures that matter for credit risk validation. All metrics are computed on D_{te}, which neither the generators nor the gate have seen.

Fidelity. The shape score averages 1 − KS over numerical attributes and 1 − total variation over categorical ones. The pair score measures agreement of association matrices (Spearman correlation, correlation ratio and Cramér's V). The detection AUC is the cross-validated AUC of a LightGBM classifier separating real from synthetic records (0.5 is ideal). Precision and recall follow the k-nearest-neighbor manifold estimator of [23] with k = 5.

Utility and model-risk agreement. A LightGBM classifier [25] (300 trees) and a logistic regression are trained on S and tested on D_{te}. We report AUC, average precision (AP), the Kolmogorov–Smirnov statistic that credit scorecards are routinely validated with, and the Brier score. Because validators examine what drives a model as well as how it ranks [10], [26], we also report the Spearman correlation ρ_{imp} between gain-based feature importances of the TSTR model and the model trained on real data (TRTR).

Privacy. The DCR share is the probability that a synthetic record is closer to a random subsample of D_{tr} than to an equally sized D_{te} (0.5 is ideal), and the near-copy rate (NCR) is the fraction of S whose DCR does not exceed the 1% quantile of test-to-train distances (0.01 is ideal). Membership inference uses random training records as members and half of D_{te} as non-members; the other half serves as the attacker's reference sample. We run three attacks: a DCR attack scoring each target by $-d(x, S)$, a DOMIAS-style density-ratio attack [19] with k-nearest-neighbor density estimates (k = 5), and a classifier-based density-ratio attack. Following [18], we report AUC and the true-positive rate at 1% false-positive rate (TPR@1%).

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

4. Experimental Setup

4.1. Datasets

We use three public credit risk datasets (Table I). Taiwan [27] contains 30,000 credit card accounts with demographic attributes, six months of repayment status, bill and payment amounts, and a default flag for the following month. Give Me Some Credit (GMSC) [28] contains 150,000 borrowers with ten numerical attributes and a two-year serious-delinquency flag; we add an indicator for missing monthly income and impute missing values with training-split medians. German [29] contains 1,000 loan applicants with 20 attributes and is included as a small-data stress test. To fit the single-core CPU budget, Taiwan and GMSC are reduced to fixed stratified subsamples of 15,000 records. Each dataset is split into 60% training, 10% calibration and 30% test records by stratified sampling; seeds 0, 1 and 2 change the split, the generator initialization and all sampling. We report means and standard deviations over the completed seeds (Taiwan: 3, GMSC: 1, German: 3).

Table I. Datasets and Splits

Dataset	Records	Num.	Cat.	Bad rate	Train/Cal/Test
Taiwan	15,000†	14	9	22.1%	9,000/1,500/4,500
GMSC	15,000†	10	1	6.7%	9,000/1,500/4,500
German	1,000	7	13	30.0%	600/100/300

†Fixed stratified subsample of the original 30,000 (Taiwan) and 150,000 (GMSC) records.

4.2. Generators and Hyperparameters

We compare five generator families. (i) A SMOTE-style interpolation synthesizer [30] with k = 5 same-class neighbors, where each categorical value is copied from the nearer endpoint. (ii) A Gaussian copula with empirical marginals and frequency-interval encoding of categories, following the Synthetic Data Vault formulation [21]. (iii) TVAE and (iv) CTGAN [20] from the authors' reference implementation (ctgan 0.12.1, batch size 500), trained for 300 epochs on German and 100 epochs on the larger datasets because of the CPU budget. (v) Our conditional diffusion model, an MLP denoiser with three hidden layers of width 256 trained with AdamW (learning rate 10^{-3}, batch size 512, exponential moving average 0.999) for 4,000 steps on German and 8,000 steps on Taiwan and GMSC. Every generator produces a candidate pool with the training class counts multiplied by r = 5 (German) or r = 3 (Taiwan, GMSC); for the cheap SMOTE and copula generators r = 6. A plain release is a uniform subsample of the pool with the training class counts; a gated release (suffix +G) applies HPQ and HBP to the same pool. PG-Diff denotes the gated diffusion release.

All experiments ran on a single CPU core with 3 GB of memory, using PyTorch 2.14, LightGBM 4.7 and scikit-learn 1.8. Code, configuration files and raw per-seed results were retained so that every number in Section V can be regenerated.

5. Results

5.1. Utility and Fidelity

Table II summarizes utility and fidelity. Models trained on real data reach AUCs of 0.762, 0.834, 0.778 on Taiwan, GMSC and German. Conditional diffusion is the most useful non-interpolating generator on two of the three datasets and is within 0.003 AUC of the best one elsewhere (GMSC: TVAE); its gap to the real-data model is 0.011, -0.011, 0.030 AUC on the three datasets. On GMSC, the TVAE and diffusion releases match or exceed the real-data model, echoing the observation of Zuo et al. [14] that synthetic data can act as a regularizer on this heavily imbalanced dataset.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

CTGAN is the weakest deep generator in our CPU-bounded setting and fails on the 600-record German split (AUC 0.520), consistent with the training instability of adversarial synthesizers on small mixed-type tables.

Fidelity metrics show why marginal agreement is a poor proxy for usefulness. The Gaussian copula has the best shape scores on Taiwan and GMSC (0.984 and 0.987) but is the least useful generator on GMSC and among the least useful on Taiwan, because its Gaussian dependence structure distorts the interactions between repayment status and amounts; a real-versus-synthetic classifier separates its records almost perfectly on Taiwan (detection AUC 0.991). Diffusion has the lowest detection AUC among the deep generators on every dataset and preserves the feature-importance ranking of the real-data model at least as well as TVAE and CTGAN on all three datasets (ρ_{imp} = 0.72, 0.95, 0.80). From a validation standpoint this matters as much as AUC: a challenger model fitted on synthetic data is only informative if it relies on the same risk drivers as the production model [26]. Overall, the ranking agrees with the comparative assessment in [4], which places diffusion ahead of GAN- and VAE-based synthesizers on fidelity and TSTR efficacy.

Table II. TSTR Utility, Fidelity and Model-Risk Agreement

Method	AUC	KS	Shape	Pair	Det.↓	ρ_{imp}
Taiwan (3 seeds)						
Real (TRTR)	0.762±0.012	0.401	0.987	0.988	0.503	—
SMOTE	0.753±0.018	0.387	0.965	0.980	0.946	0.919
Copula	0.691±0.018	0.293	0.984	0.890	0.991	0.692
TVAE	0.742±0.007	0.383	0.896	0.938	0.998	0.245
CTGAN	0.699±0.020	0.303	0.805	0.900	0.999	0.446
Diffusion	0.751±0.009	0.386	0.916	0.916	0.946	0.723
PG-Diff	0.748±0.005	0.384	0.917	0.915	0.944	0.702
GMSC (1 seed)						
Real (TRTR)	0.834±0.000	0.537	0.989	0.990	0.493	—
SMOTE	0.825±0.000	0.520	0.985	0.987	0.624	0.936
Copula	0.678±0.000	0.297	0.987	0.954	0.882	0.645
TVAE	0.848±0.000	0.539	0.847	0.933	0.993	0.909
CTGAN	0.779±0.000	0.414	0.879	0.938	0.998	0.818
Diffusion	0.845±0.000	0.517	0.941	0.969	0.789	0.945
PG-Diff	0.843±0.000	0.532	0.941	0.967	0.760	0.973
German (3 seeds)						
Real (TRTR)	0.778±0.024	0.443	0.967	0.957	0.486	—
SMOTE	0.754±0.041	0.401	0.949	0.953	0.717	0.927
Copula	0.613±0.062	0.196	0.962	0.923	0.609	0.878
TVAE	0.724±0.036	0.392	0.769	0.887	0.977	0.641
CTGAN	0.520±0.028	0.090	0.910	0.911	0.886	0.781
Diffusion	0.748±0.023	0.411	0.937	0.942	0.656	0.797
PG-Diff	0.735±0.025	0.395	0.943	0.941	0.643	0.879

AUC: mean±std over seeds; other columns: means. Det.: detection AUC of a real-vs-synthetic classifier (0.5 ideal). Best synthetic value per dataset in bold. PG-Diff = diffusion + HPQ + HBP gate.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

5.2. Privacy Audit

Table III reports the privacy audit. The SMOTE synthesizer leaks on every dataset: its DCR share is 0.736–0.773, its near-copy rate is 0.23–0.92 against an ideal of 0.01, and the distance attack identifies members with an AUC of 0.739–0.763 and a TPR of 0.19–0.58 at 1% false positives. Its competitive utility in Table II is therefore bought with disclosure risk, in line with the warning of Stadler et al. [5] that synthetic data are not automatically anonymous. Under the three attacks we ran, the other generators sit close to the ideal values on Taiwan and GMSC, where all distance-attack AUCs fall in 0.477–0.504. On the 600-record German split, plain diffusion shows a mild excess proximity (DCR share 0.549) that the attacks only partly convert into membership signal. The DOMIAS-style and classifier attacks are weaker than the distance attack against interpolated records, which illustrates why a privacy audit should combine several attack families rather than rely on a single score [15].

Table III. Privacy Audit on Held-Out Records

Method	DCR sh.	NCR	MIA-DCR	TPR@1%	DOMIAS	Clf.
Taiwan						
SMOTE	0.773	0.259	0.739±0.004	0.201	0.525	0.510
SMOTE+G	0.752	0.045	0.704±0.007	0.071	0.509	0.504
Copula	0.510	0.000	0.501±0.007	0.009	0.501	0.499
TVAE	0.515	0.000	0.504±0.004	0.011	0.500	0.505
CTGAN	0.497	0.000	0.503±0.009	0.011	0.502	0.507
Diffusion	0.504	0.003	0.504±0.010	0.010	0.504	0.509
PG-Diff	0.506	0.003	0.502±0.005	0.008	0.505	0.511
GMSC						
SMOTE	0.767	0.225	0.763±0.000	0.190	0.534	0.533
SMOTE+G	0.724	0.107	0.717±0.000	0.115	0.515	0.523
Copula	0.489	0.000	0.487±0.000	0.016	0.503	0.500
TVAE	0.469	0.000	0.477±0.000	0.012	0.481	0.497
CTGAN	0.483	0.000	0.486±0.000	0.010	0.504	0.494
Diffusion	0.503	0.002	0.504±0.000	0.009	0.503	0.499
PG-Diff	0.500	0.002	0.500±0.000	0.006	0.496	0.498
German						
SMOTE	0.736	0.924	0.746±0.005	0.580	0.510	0.601
SMOTE+G	0.714	0.680	0.695±0.073	0.444	0.518	0.566
Copula	0.511	0.001	0.476±0.039	0.011	0.475	0.528
TVAE	0.502	0.084	0.486±0.019	0.020	0.487	0.511
CTGAN	0.519	0.000	0.487±0.012	0.011	0.500	0.524
Diffusion	0.549	0.013	0.503±0.033	0.022	0.514	0.549
PG-Diff	0.536	0.007	0.484±0.021	0.009	0.506	0.556

Ideal values: DCR share 0.5, NCR 0.01, attack AUCs 0.5, TPR@1% 0.01. MIA-DCR: distance attack (mean±std); DOMIAS: k-NN density-ratio attack; Clf.: classifier density-ratio attack.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

5.3. Effect of the Release Gate

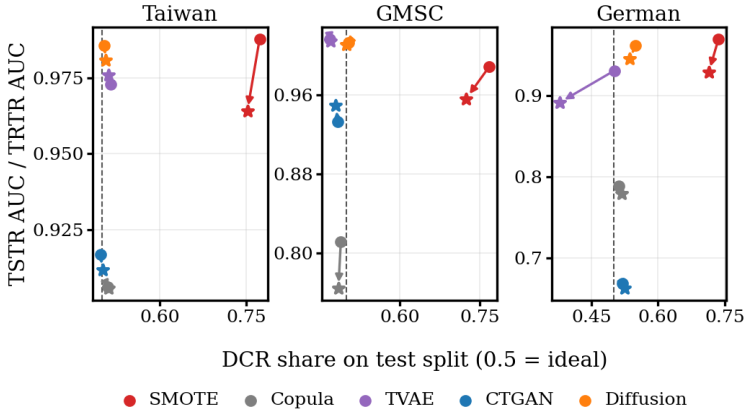


Figure 2. Privacy–utility effect of the release gate. Circles: plain releases; stars: gated releases (+G) from the same candidate pool; arrows show the change. The x-axis is the DCR share on the test split (0.5 is ideal); the y-axis is TSTR AUC relative to the real-data model.

Fig. 2 and Table IV show what the gate does to each generator. For the four generators that do not interpolate between training records, the gate changes TSTR AUC by between -0.039 and 0.013 and keeps the measured privacy indicators close to their ideal values, so it acts as an inexpensive check against the evaluated empirical privacy criteria. For PG-Diff in particular, the AUC difference to plain diffusion is -0.004, -0.002, -0.013 on Taiwan, GMSC and German. For SMOTE the gate cannot repair the release, but it does two useful things. First, it removes the most exposed records: the TPR at 1% false positives falls from 0.201, 0.190, 0.580 to 0.071, 0.115, 0.444. Second, it reports that admissible candidates ran out: 13% on Taiwan, 31% on GMSC, 97% on German of the gated SMOTE release had to be filled by relaxation, even from a pool six times larger than the release. We read a high relaxation rate as evidence that the generator does not satisfy the evaluated empirical privacy criteria, a signal that is more informative than any single privacy score. On German the gate shifted the TVAE release below the ideal DCR share (0.378) because its candidates concentrate on a few modes (recall 0.03); a two-sided balance constraint addresses this case.

Table IV. Effect of the Release Gate on Each Generator

Generator	ΔAUC	DCR share	MIA-DCR AUC	TPR@1%	Relaxed
SMOTE	-0.025	0.759→0.730	0.749→0.705	0.324→0.210	47.0%
Copula	-0.016	0.503→0.505	0.488→0.483	0.012→0.008	0.0%
TVAE	-0.010	0.495→0.453	0.489→0.490	0.014→0.009	5.5%
CTGAN	+0.002	0.500→0.501	0.492→0.494	0.011→0.010	0.0%
Diffusion	-0.006	0.519→0.514	0.504→0.495	0.014→0.008	0.0%

Plain → gated values averaged over the three datasets and all seeds. Relaxed: share of the gated release that had to be filled after the admissible candidates were exhausted.

5.4. Ablation and Memorization Stress Test

Table V isolates the two constraints. Under the default training budget, HPQ and HBP are individually and jointly almost neutral on Taiwan and GMSC, which is the expected behavior when the generator does not memorize. To test the failure mode documented for tabular diffusion [9], we retrained the German model four times longer. Extended training raised the DCR share of plain diffusion from 0.549 to 0.595 and the near-copy rate from 0.013 to 0.023, while lowering TSTR AUC from 0.748 to 0.709, a clear sign of overfitting. PG-Diff returned the DCR share to 0.472 and the near-copy rate to 0.012 at a cost of 0.033 AUC. Fig. 3 shows the mechanism: the plain release has an excess of records in the lower tail of the DCR distribution

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

relative to fresh test records, and the gated release follows the fresh-data profile. Fidelity importance resampling (FIR) is a useful optional step when the calibration split is large enough: it raised the shape score on Taiwan from 0.917 to 0.962 and lowered the GMSC detection AUC from 0.760 to 0.659 with an AUC change of 0.000 and -0.002, respectively.

Table V. Ablation of the Gate and Memorization Stress Test (AUC / DCR Share)

Variant	Taiwan	GMSC	German	German-long
Diffusion (no gate)	0.751 / 0.504	0.845 / 0.503	0.748 / 0.549	0.709 / 0.595
+ HPQ only	0.748 / 0.507	0.843 / 0.500	0.735 / 0.541	0.698 / 0.478
+ HBP only	0.746 / 0.504	0.843 / 0.500	0.734 / 0.548	0.671 / 0.493
+ HPQ + HBP (PG-Diff)	0.748 / 0.506	0.843 / 0.500	0.735 / 0.536	0.676 / 0.472
+ HPQ + HBP + FIR	0.748 / 0.507	0.841 / 0.498	0.746 / 0.543	—

German-long: diffusion trained for 16,000 instead of 4,000 steps (1 seed); FIR was not run in the stress test.

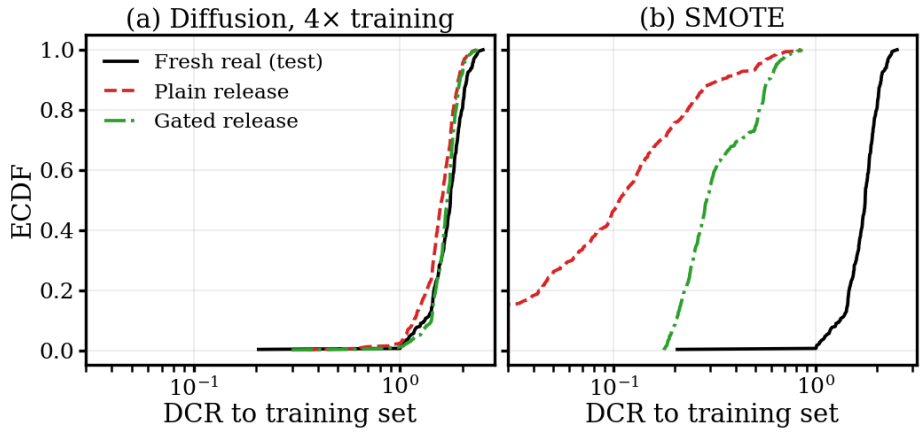


Figure 3. Empirical CDF of the distance to the closest training record on German (seed 0). (a) Diffusion trained four times longer; (b) SMOTE. A curve above the fresh-data curve at small distances indicates excess proximity to training records.

5.5. Guidance Scale and Cost

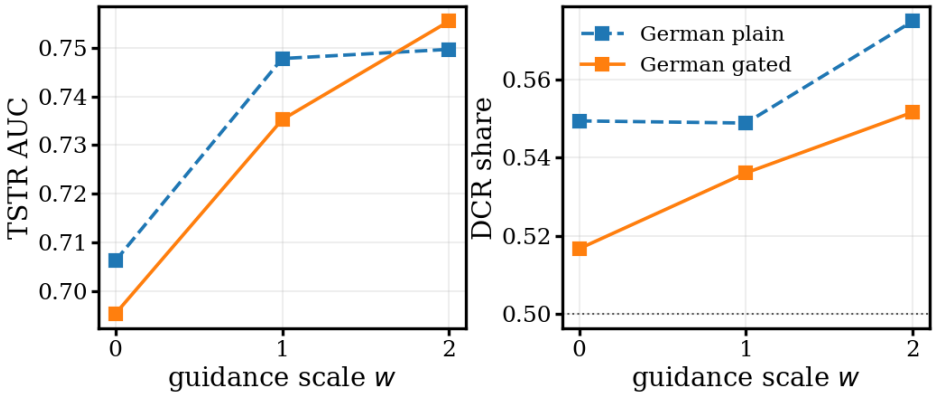


Figure 4. Effect of the classifier-free guidance scale w on TSTR AUC and DCR share for plain (dashed) and gated (solid) diffusion releases.

Guidance strengthens the class signal that PD models learn from. On German, moving from w = 0 to w = 1 raised TSTR AUC from 0.706 to 0.748 (Fig. 4). Stronger guidance (w = 2) changed AUC by 0.002 but increased the German DCR share to 0.575, which the gate reduced to 0.552. Guidance therefore trades diversity for separation, and the gate is a useful check when it is pushed. Computationally, the gate is cheap: on Taiwan it took 4.1 s per release on a single CPU

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

core, compared with 2.5 min to train the diffusion model, 0.8 min for TVAE and 3.2 min for CTGAN.

6. Discussion: Applications, Challenges and Future Directions

6.1. Applications in Credit Risk Management

The results suggest concrete roles for generative AI in a credit risk function. Gated diffusion releases satisfying the evaluated empirical privacy criteria can support model development and benchmarking by external teams, since PD models trained on them rank real borrowers almost as well as models trained on the source data and rely on similar drivers. They can serve challenger-model and validation exercises required by model risk guidance [10], where a reproducible synthetic copy of a development sample allows independent re-estimation without sharing the original records. Because the generator is conditioned on the default label, the same machinery can produce stressed portfolios with a higher default share or targeted minority samples, which connects synthetic data to the scenario-generation use cases surveyed in [1]. Finally, the admissibility rate of the gate gives risk and privacy officers a simple, auditable screening signal: a generator whose candidates cannot pass the fresh-data test fails the evaluated empirical privacy criteria, whatever its utility, whereas passing the test is necessary but not sufficient evidence of low disclosure risk. Temporal Transformer and conditional tabular generation experiments in financial fraud detection demonstrate the value of sequence-sensitive learning together with controlled synthetic-data augmentation [32].

6.2. Open Challenges

Several challenges remain open. The gate is an empirical safeguard that reads the training data, so repeated releases or interactive access to privacy metrics should be governed to prevent the reconstruction attacks described by Ganev and De Cristofaro [15], and the gate is best combined with DP training where formal guarantees are required. Extending the audit from distance-based, density-ratio and classifier-based membership inference to shadow-model and attribute-inference attacks, and scaling the generators beyond the single-CPU budget used here, are natural next steps.

6.3. Future Directions

Four directions follow. First, DP-trained diffusion models combined with holdout-calibrated gates could combine formal guarantees with an empirical audit of the released records. Second, a two-sided balance constraint and adaptive pool sizes would remove the over-correction and relaxation effects observed here. Third, extending the approach to behavioral and transactional data, where temporal dependence carries much of the risk signal [2], is essential for fraud and early-warning models. Fourth, large language models [3] are beginning to act as tabular generators and as agents that consume synthetic data; release gates and privacy audits will need to cover these pipelines as well, together with fairness-aware conditioning so that synthetic portfolios do not reproduce historical lending bias.

7. Conclusion

This paper examined generative AI for credit risk data through the lens of a release decision: how useful a synthetic portfolio is for risk modeling, and whether it satisfies the evaluated empirical privacy criteria. We proposed PG-Diff, which combines a risk-conditional tabular diffusion model with a release gate calibrated on held-out real records, and evaluated it with a three-axis fidelity–utility–privacy protocol extended with low-false-positive membership inference and model-risk agreement. On three public credit datasets, conditional diffusion produced the most realistic records among the deep generators, with TSTR AUCs within 0.030

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

of real-data models. The gate was nearly free for well-behaved generators, removed the excess proximity created by overtraining, and showed that interpolation-based synthesis fails the evaluated empirical privacy criteria. Holdout-calibrated release gates reduced measured record-proximity and membership-inference risks in all evaluated settings. We therefore see them as a simple, generator-agnostic and auditable building block for governing generative AI in financial risk management, to be combined with formal mechanisms where guarantees are required.

References

[1] Assefa, S. A., Dervovic, D., Mahfouz, M., Tillman, R. E., Reddy, P., & Veloso, M. (2020). Generating synthetic data in finance: Opportunities, challenges and pitfalls. In Proceedings of the 1st ACM International Conference on AI in Finance (ICAIF). https://doi.org/10.1145/3383455.3422554

[2] Wiese, M., Knobloch, R., Korn, R., & Kretschmer, P. (2020). Quant GANs: Deep generation of financial time series. Quantitative Finance, 20(9), 1419–1440. https://doi.org/10.1080/14697688.2020.1730426

[3] Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A large language model for finance. arXiv:2303.17564.

[4] Yang, Y., & Yang, J. (2026). Synthetic data meets finance: Generative models for privacy preserving analytics. Journal of Banking Financial Dynamics, 10(4), 1–8. https://doi.org/10.55220/2576-6821.v10.928

[5] Stadler, T., Oprisanu, B., & Troncoso, C. (2022). Synthetic data—Anonymisation groundhog day. In Proceedings of the 31st USENIX Security Symposium (pp. 1451–1468).

[6] Kotelnikov, A., Baranchuk, D., Rubachev, I., & Babenko, A. (2023). TabDDPM: Modelling tabular data with diffusion models. In Proceedings of the 40th International Conference on Machine Learning (ICML) (Vol. 202, pp. 17564–17579). PMLR.

[7] Zhang, H., Zhang, J., Srinivasan, B., Shen, Z., Qin, X., Faloutsos, C., Rangwala, H., & Karypis, G. (2024). Mixed-type tabular data synthesis with score-based diffusion in latent space. In Proceedings of the International Conference on Learning Representations (ICLR).

[8] Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., & Wallace, E. (2023). Extracting training data from diffusion models. In Proceedings of the 32nd USENIX Security Symposium (pp. 5253–5270).

[9] Chen, J., Zhang, X., & Han, X. (2023). Explainable identification of tax-related financial reporting anomalies in U.S. public companies: Cross-fitted tax-consistency residuals and Owen-value Shapley propagation. World Journal of Economics and Business Research, 1(1), 33–45.

[10] Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency. (2011, April). Supervisory guidance on model risk management (SR Letter 11-7).

[11] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407. https://doi.org/10.1561/0400000042

[12] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS) (pp. 308–318). https://doi.org/10.1145/2976749.2978318

[13] Jordon, J., Yoon, J., & van der Schaar, M. (2019). PATE-GAN: Generating synthetic data with differential privacy guarantees. In Proceedings of the International Conference on Learning Representations (ICLR).

[14] Zuo, M., Kang, I., Patterson, S., & Seneviratne, O. (2026). Measuring privacy risks and tradeoffs in financial synthetic data generation. In Companion Proceedings of the ACM Web Conference (WWW Companion). https://doi.org/10.1145/3774905.3797244

[15] Ganev, G., & De Cristofaro, E. (2025). The inadequacy of similarity-based privacy metrics: Privacy attacks against "truly anonymous" synthetic datasets. In Proceedings of the 46th IEEE Symposium on Security and Privacy (SP).

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

[16] Platzer, M., & Reutterer, T. (2021). Holdout-based empirical assessment of mixed-type synthetic data. Frontiers in Big Data, 4, 679939. https://doi.org/10.3389/fdata.2021.679939

[17] Ho, J., & Salimans, T. (2022). Classifier-free diffusion guidance. arXiv:2207.12598.

[18] Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., & Tramèr, F. (2022). Membership inference attacks from first principles. In Proceedings of the IEEE Symposium on Security and Privacy (SP) (pp. 1897–1914). https://doi.org/10.1109/SP46214.2022.9833649

[19] Yue, X., & Zhu, R. (2024). An explainable hybrid machine learning framework for cash-flow-conditioned multi-horizon liquidity risk early warning: An SME-oriented benchmark study. World Journal of Management Science, 2(1), 63–72.

[20] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. In Advances in Neural Information Processing Systems (Vol. 32).

[21] Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. In Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 399–410). https://doi.org/10.1109/DSAA.2016.49

[22] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (Vol. 33, pp. 6840–6851).

[23] Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., & Aila, T. (2019). Improved precision and recall metric for assessing generative models. In Advances in Neural Information Processing Systems (Vol. 32).

[24] Song, J., Meng, C., & Ermon, S. (2021). Denoising diffusion implicit models. In Proceedings of the International Conference on Learning Representations (ICLR).

[25] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems (Vol. 30, pp. 3146–3154).

[26] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. Computational Economics, 57(1), 203–216. https://doi.org/10.1007/s10614-020-10042-0

[27] Yeh, I.-C., & Lien, C.-H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020

[28] Teng, D. (2025). TRIAD: Deadline- and reliability-aware scheduling for post-earthquake structural assessment workflows. World Journal of Engineering Research, 3(6), 55–66.

[29] Hofmann, H. (1994). Statlog (German Credit Data). UCI Machine Learning Repository. https://doi.org/10.24432/C5NC77

[30] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[31] Han, X., et al. (2025). Symmetry-Aware Credit Risk Modeling: A Deep Learning Framework Exploiting Financial Data Balance and Invariance. Symmetry, 17(3), 341. https://doi.org/10.3390/sym17030341

[32] Chen, J., et al. (2026). Temporal Transformer with Conditional Tabular GAN for Credit Card Fraud Detection: A Sequential Deep Learning Approach. Mathematics, 14, 1183. https://doi.org/10.3390/math14071183