BOLT: BOOLEAN-FEATURE ONE-RELEASE LABEL TRANSFER FOR DIFFERENTIALLY PRIVATE TABULAR SYNTHETIC DATA GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Differentially private tabular synthesis must balance predictive utility with the computational cost of producing usable data. We introduce BOLT (Boolean-feature One-release Label Transfer), which constructs two complementary models from a single jointly calibrated Gaussian release: a simple class-conditional covariate sampler and a richer Bernoulli naïve-Bayes classifier over fixed threshold predicates and randomly selected two-literal conjunctions. BOLT samples covariates from the simpler model and assigns stochastic labels using the classifier, incorporating private interaction statistics into prediction without fitting a generator to reproduce those statistics. Under a fixed public schema and encoding, synthesis is postprocessing of the shared (ε, δ)-differentially private release. We evaluate five private synthesis methods across five datasets, four privacy budgets ($\varepsilon \in \{0.5, 1, 2, 8\}$), and two downstream classifiers. Averaging F1 and AUROC ratios to the corresponding nonprivate real-data references equally across datasets, budgets, learners, and metrics, BOLT retains 87.0% of reference utility with a mean complete-table synthesis time of 2.85 seconds, compared with 89.3% and 5.16 hours for AIM. BOLT achieves the lowest mean synthesis time in all twenty dataset–privacy configurations. Ratios of recorded arithmetic mean runtimes give $12.9\times$ and $6{,}517\times$ speed advantages over DP-MERF, the next-fastest method, and AIM, respectively. After averaging scores across releases, BOLT exceeds each of DP-MERF, DP-HP, and Tab-PE in at least 80% of the dataset–budget–learner–metric comparisons.

1 INTRODUCTION

For a fixed prediction task, a synthetic table needs to preserve information useful for learning the target. Private set generation has already pursued this goal by directly optimizing synthetic samples using discriminative information rather than fitting a full generative distribution (Chen et al., 2022). We ask whether private interaction measurements can instead construct useful labels while keeping covariate generation simple.

BOLT pairs two models with different representations. A mixture of product distributions generates covariates from class-indexed univariate statistics. A Bernoulli naïve-Bayes classifier combines numerical thresholds and two-literal conjunctions to determine their labels. A single jointly calibrated Gaussian summary supplies both models, allowing interaction measurements to influence prediction without fitting a generator to reproduce them.

We evaluate whether this separation reduces complete-table synthesis time while retaining predictive utility. Across five tasks and two downstream learners, BOLT achieves the lowest mean synthesis time in every evaluated dataset–privacy configuration, with dataset-averaged utility usually second to AIM. Component experiments examine how the conjunction workload and labeling step contribute to this tradeoff.

2 BACKGROUND AND RELATED WORK

Private statistics and synthetic reconstruction. Private naïve-Bayes methods estimate class frequencies and class-conditional parameters from perturbed statistics (Cormode, 2010; Vaidya et al.,

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.

2013). Kernel-mean release and DP-MERF privatize bounded feature averages before fitting synthetic outputs or a generator (Balog et al., 2018; Harder et al., 2021). Posterior sampling offers another route to differential privacy under bounded log-likelihood assumptions (Wang et al., 2015). DP-HP and DP-NTK use Hermite and neural-tangent features, respectively, to form private summaries for subsequent generator fitting (Vinaroz et al., 2022; Yang et al., 2024).

Classifier-guided synthesis and interaction features. QUAIL is a close architectural precedent: it separately privatizes a synthesizer and a supervised model, including a Gaussian naïve-Bayes instantiation, then appends predicted labels to generated covariates (Rosenblatt et al., 2020). BOLT instead derives a class-conditional product sampler and a Bernoulli naïve-Bayes classifier over predicates and conjunctions from one jointly calibrated Gaussian summary, then draws stochastic labels subject to an explicit expected-prevalence constraint. Nonprivate model compression labels generated inputs with a teacher (Buciluǎ et al., 2006). Constructed attributes have long incorporated interactions into naïve Bayes; for example, Pazzani (1996) considers Cartesian-product attributes. That construction differs from our fixed Boolean conjunctions, but establishes the broader ancestry of interaction-enriched naïve Bayes. Our combination uses one release for both models and Bernoulli label draws with an explicit prevalence constraint.

Supervised queries and marginal models. RAP++ studies private synthetic data through query families useful for classification, including label-conditioned marginal and threshold queries (Vietri et al., 2022). A class-indexed conjunction average in our release combines a selected joint-event frequency involving two predictors and the label with the corresponding class frequency. PrivBayes samples from a Bayesian network estimated using private low-dimensional marginals (Zhang et al., 2014). Private-PGM and AIM reconstruct graphical-model distributions from noisy marginals (McKenna et al., 2019; 2022). BOLT instead uses the measured interactions in a prediction score; it does not solve a joint-distribution consistency problem.

3 Mechanism

3.1 Public representation and Boolean features

Let $D = \{(x_i, y_i)\}_{i=1}^n$ with $y_i \in \{0, 1\}$. The public encoded domain is $\mathcal{X} = [-1, 1]^d \times \prod_{k=1}^p \{1, \dots, K_k\}$. Here d counts numerical columns and p categorical columns, with $d + p > 0$. Bounds, category domains, and rowwise transformations are fixed independently of the confidential dataset. Privacy is defined at this encoded input; Section 4 specifies adjacency and the normalization constant n_0.

For each numerical coordinate, form predicates $\mathbf{1}\{x_j > t\}$ at seven fixed thresholds $t \in \{-3/4, -1/2, -1/4, 0, 1/4, 1/2, 3/4\}$. Categorical predicates are equalities $\mathbf{1}\{x_k = v\}$. A literal is a predicate b or its negation $1 - b$. A two-literal conjunction is their Boolean AND, equivalently $z(x) = \ell_a(x)\ell_b(x)$, where the two literals refer to different original columns. Public randomness selects $J = 512$ distinct conjunction specifications, independently of D. We assume the public schema permits the requested number of cross-column specifications. Smaller or larger J define the conjunction-count ablations.

The Boolean vector $z(x) \in \{0, 1\}^M$ contains the $7d$ numerical predicates and the J conjunctions, so $M = 7d + J > 0$. Standalone categorical equalities enter the categorical statistics below and may appear inside conjunctions; they are not separate terms in z. Their signed numerical encoding is $a(x) = 2z(x) - \mathbf{1}$.

Figure 1 summarizes how BOLT constructs its covariate sampler and labeling classifier from one private release.

3.2 One shared private summary

Write $v(x)$ for the concatenation of the encoded numerical coordinates and the categorical one-hot vectors. With $L = d + p$, define

$$g(x) = \frac{v(x)}{\sqrt{L}}, \qquad \psi(x) = \left(\frac{g(x)}{\sqrt{2}}, \frac{a(x)}{\sqrt{2M}}\right), \qquad h(x, y) = \frac{(e_y, e_y \otimes \psi(x))}{\sqrt{2}}, \tag{1}$$

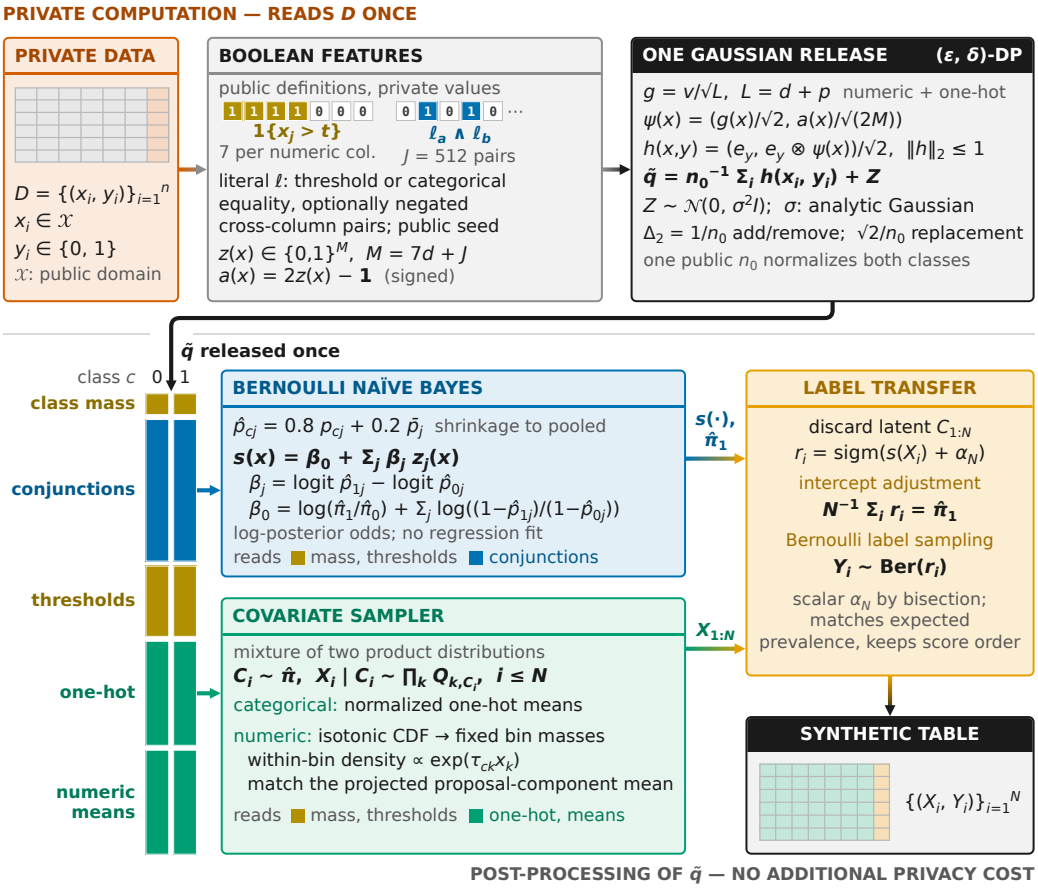

Figure 1: **BOLT architecture.** A jointly calibrated Gaussian release supplies a class-conditional covariate sampler and a Boolean-feature classifier. Sampled component labels are replaced by stochastic classifier labels after an intercept adjustment. The diagram omits the numerical clipping of the prevalence target in (8).

where e_y is the two-dimensional label indicator. Since $\|g(x)\|_2 \le 1$ and $\|a(x)\|_2^2 = M$, both $\|\psi(x)\|_2$ and $\|h(x,y)\|_2$ are at most one. The release is

$$q(D) = \frac{1}{n_0}\sum_{i=1}^{n} h(x_i, y_i), \qquad \widetilde{q} = q(D) + Z, \qquad Z \sim \mathcal{N}(0, \sigma^2 I). \tag{2}$$

The same n_0 normalizes every class block. Thus q contains class-indexed sums divided by n_0, rather than separately normalized conditional means. The noise is calibrated jointly for the whole vector.

3.3 Bernoulli naïve Bayes from noisy statistics

Let $\widetilde{q}_c$ be a class-mass coordinate and $\widetilde{q}_{\psi,c}$ its feature block. Define

$$m_c = \max\{\sqrt{2}\widetilde{q}_c, 10^{-6}\}, \qquad \widehat{\pi}_c = \frac{m_c}{m_0 + m_1}, \qquad \widehat{\mu}_c = \frac{\sqrt{2}\widetilde{q}_{\psi,c}}{m_c}. \tag{3}$$

The denominator in $\widehat{\mu}_c$ is the unnormalized floored mass m_c, not $\widehat{\pi}_c$. For a Boolean coordinate j, undo the scaling in (1) and convert its signed mean to a probability:

$$u_{cj} = \mathrm{clip}\big(\sqrt{2M}\,\widehat{\mu}_{c,a,j}, -1, 1\big), \quad p_{cj} = \mathrm{clip}\big((1+u_{cj})/2, 10^{-3}, 1-10^{-3}\big). \tag{4}$$

Here $\widehat{\mu}_{c,a,j}$ denotes a Boolean coordinate in the second block of ψ; $\widehat{\mu}_{c,g,k}$ denotes numerical coordinate k in the first block. With fixed shrinkage $\lambda = 0.2$, set

$$\bar{p}_j = \sum_{c=0}^{1} \widehat{\pi}_c p_{cj}, \qquad \widehat{p}_{cj} = (1-\lambda)p_{cj} + \lambda \bar{p}_j. \tag{5}$$

The Bernoulli naïve-Bayes log-posterior odds are

$$s(x) = \log \frac{\widehat{\pi}_1}{\widehat{\pi}_0} + \sum_{j=1}^{M} \left[z_j(x) \log \frac{\widehat{p}_{1j}}{\widehat{p}_{0j}} + (1 - z_j(x)) \log \frac{1-\widehat{p}_{1j}}{1-\widehat{p}_{0j}} \right]. \tag{6}$$

Both satisfied and unsatisfied predicates contribute. Equivalently, $s(x) = \beta_0 + \sum_j \beta_j z_j(x)$, where $\beta_j = \mathrm{logit}(\widehat{p}_{1j}) - \mathrm{logit}(\widehat{p}_{0j})$. These coefficients are computed directly from probability contrasts.

The fitted-model probability is $\mathrm{sigmoid}(s(x))$, where $\mathrm{sigmoid}(t) = (1 + e^{-t})^{-1}$. Conjunctions introduce pairwise Boolean interactions in the log odds. The conditional-independence assumption is a working model for this expanded feature vector.

3.4 Covariates from a simpler model

For each class c, construct a product distribution $Q_c = \prod_k Q_{kc}$ from univariate statistics in the same $\widetilde{q}$. Categorical probabilities use the corresponding one-hot means after flooring and normalization. For numerical columns, threshold probabilities define a noisy cumulative distribution function. Isotonic regression makes it nondecreasing; differences give bin probabilities, followed by positive flooring and normalization. Holding these probabilities fixed, sample within each bin with density proportional to $\exp(\tau_{ck} x_k)$. One scalar τ_{ck} per class and column matches the released encoded mean $\sqrt{2L}\,\widehat{\mu}_{c,g,k}$ after projection onto the feasible interval. Zero tilt gives uniform sampling; boundary targets use endpoint limits. Appendix B specifies the fit and numerical handling.

Draw a latent component $C_i \sim \widehat{\pi}$ and then $X_i \sim Q_{C_i}$, independently across rows. The covariate distribution is

$$Q_X(dx) = \sum_{c=0}^{1} \widehat{\pi}_c \prod_k Q_{kc}(dx_k). \tag{7}$$

Columns are independent conditional on C_i, but can be dependent marginally. Numerical means shape values within bins; conjunctions enter the classifier, not the proposal. The mean constraints concern proposal components C_i, not the final labels Y_i.

3.5 Stochastic labels and expected prevalence

Discard the sampled component labels $C_{1:N}$. Set $\rho_* = \mathrm{clip}(\widehat{\pi}_1, 10^{-9}, 1 - 10^{-9})$ and choose the scalar α_N satisfying

$$\frac{1}{N} \sum_{i=1}^{N} \mathrm{sigmoid}\big(s(X_i) + \alpha_N\big) = \rho_*. \tag{8}$$

The main configuration uses temperature one. The left side is continuous and strictly increasing from zero to one, so the offset exists and is unique. The implementation finds it by bisection. Finally, independently conditional on $X_{1:N}$ and $\widetilde{q}$, draw

$$Y_i \sim \mathrm{Bernoulli}\big(r_i\big), \qquad r_i = \mathrm{sigmoid}\big(s(X_i) + \alpha_N\big). \tag{9}$$

The table has N equally weighted rows; its realized positive count remains random.

Proposition 1 (Prevalence adjustment)**.** *For finite scores $s_i = s(X_i)$ and $0 < \rho_* < 1$, the probabilities in (9) uniquely minimize*

$$\sum_{i=1}^{N} D_{\mathrm{KL}}(\mathrm{Bernoulli}(r_i) \,\|\, \mathrm{Bernoulli}(\mathrm{sigmoid}(s_i))) \quad \textit{subject to} \quad 0 \le r_i \le 1, \quad \frac{1}{N}\sum_i r_i = \rho_*. \tag{10}$$

They preserve score ordering and satisfy $\mathbb{E}[N^{-1}\sum_i Y_i \mid X_{1:N}, \widetilde{q}] = \rho_*$.

This is a constrained information projection (Csiszár, 1975) and a special case of expectation-constrained posterior adjustment (Ganchev et al., 2010); Appendix A gives the proof. Because α_N uses the whole covariate batch, complete output rows are generally not independent conditional on the release alone.

4 Privacy analysis

A randomized mechanism $\mathcal{A}$ is (ε, δ)-differentially private if, for all neighboring D, D' and measurable output events S, $\mathbb{P}[\mathcal{A}(D) \in S] \leq e^{\varepsilon}\mathbb{P}[\mathcal{A}(D') \in S] + \delta$ (Dwork et al., 2006; Dwork & Roth, 2014). For add/remove adjacency, neighbors differ by one record and use the same public denominator $n_0 > 0$. For replacement adjacency, both contain a fixed public number $n = n_0$ of records. The public schema, Boolean-feature construction, output size N, privacy parameters, and numerical constants are fixed across neighbors. In particular, add/remove privacy does not permit substituting the observed sample size for n_0 or revealing that size through N.

Theorem 2 (Privacy of the shared-summary mechanism)**.** *Fix $\varepsilon > 0$ and $0 < \delta < 1$. On the public domain of Section 3.1, the global L_2 sensitivity of (2) is $\Delta = 1/n_0$ under add/remove adjacency and $\Delta = \sqrt{2}/n_0$ under replacement adjacency. Let $\sigma > 0$ satisfy*

$$\Phi\left(\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) - e^{\varepsilon}\Phi\left(-\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) \leq \delta, \tag{11}$$

where Φ is the standard-normal CDF. Then $\widetilde{q}$, the two fitted models, and any fixed finite collection of tables generated from this same release jointly satisfy (ε, δ)-DP.

The proof combines the exact sensitivity of this feature map with the analytic Gaussian mechanism of Balle & Wang (2018) and the post-processing property. A single record has norm at most one; the largest replacement distance is $\sqrt{2}$, attained by changing the label of a norm-one record. Both models, including numerical-mean projection and within-bin tilt fitting, and the entire table use only $\widetilde{q}$, public inputs, and fresh randomness independent of the data and the release noise. Appendix A supplies the full argument, including both same-label and different-label replacements.

Independently recomputed releases from the same confidential table require privacy composition. The theorem assumes a fixed public representation and does not cover private-data-dependent evaluation or runtime measurements.

5 Experimental Procedure

We test whether BOLT can reduce the time required to produce a private synthetic table while retaining competitive predictive utility. The primary comparison therefore measures downstream classification performance alongside the time to the first complete synthetic table.

Tasks and datasets. The core benchmark comprises five binary classification tasks: income prediction (Adult), credit-card default (Default), hospital readmission (Diabetes), gamma/hadron discrimination (MAGIC), and purchase prediction (Shoppers). We retain fixed public training/test partitions, spanning 11,097–61,059 training rows and 10–36 predictors. All methods receive the same encoded inputs: numerical variables mapped into $[-1, 1]$ using fixed public transformations and categorical variables assigned fixed domains with an unknown category. Category domains come from the public benchmark snapshot.

Methods and privacy. We compare BOLT with schema-adapted configurations of AIM (McKenna et al., 2022), DP-MERF (Harder et al., 2021), DP-HP (Vinaroz et al., 2022), and Tab-PE (Tran et al., 2026), covering marginal reconstruction, private feature matching with neural generators, and iterative private selection and mutation. Each method produces $N = n_0$ rows, where n_0 is the public training-set reference size. The main grid uses add/remove adjacency, $\delta = n_0^{-2}$, and $\varepsilon \in \{0.5, 1, 2, 8\}$. Each configuration has five independent private releases, except AIM, which has three, yielding 460 synthetic tables. Configurations are shared across datasets without final-test tuning; baseline adaptations and exact settings appear in Appendix C.

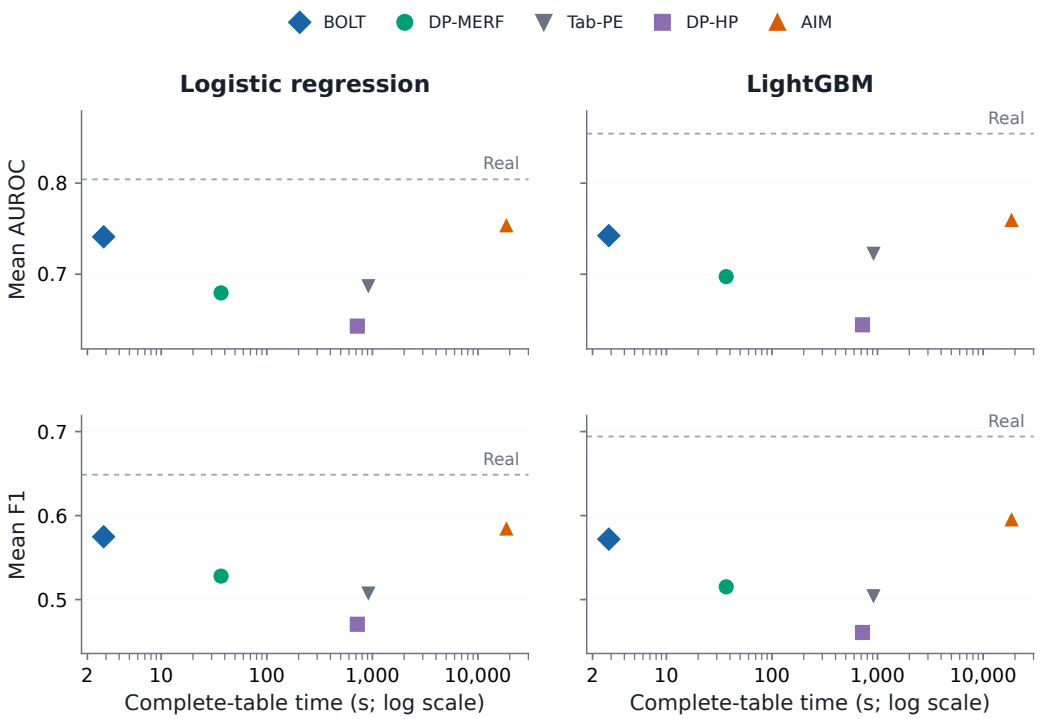

Figure 2: **Predictive utility and complete-table synthesis time.** Points summarize the same twenty dataset–privacy configurations as the headline tables. Release scores and runtimes are averaged within each configuration, then configurations receive equal weight. Higher utility and shorter time are preferable; runtime axes are logarithmic. Dashed lines show the nonprivate Real utility reference.

Predictive utility. We train logistic regression and LightGBM (Ke et al., 2017) separately on each synthetic table, testing linear and nonlinear transfer to the original real test set. Both learners receive numerical values and categorical one-hot vectors, without BOLT's Boolean feature expansion. A stratified 80/20 synthetic split separates model fitting from threshold selection. F1 uses the threshold maximizing synthetic-holdout F1 over $\{0.02, 0.03, \ldots, 0.98\}$; AUROC uses predicted probabilities. Real test labels are used only for final scoring. A nonprivate Real reference follows the corresponding real-training fit/holdout procedure. We average releases within each configuration, then weight datasets equally; summaries across privacy budgets also weight budgets equally.

Component tests. At $\varepsilon = 1$, five releases per dataset compare 0, 128, 512, and 2,048 conjunctions from nested public collections. Each workload is calibrated to the same privacy budget, so this comparison includes changes in effective measurement noise. A separate paired experiment holds the private summary and generated covariates fixed, comparing BOLT's stochastic classifier labels with the latent component labels used to sample those covariates.

Runtime. Timed synthesis begins after input preparation. Complete-table latency includes private measurement, model fitting, and generation, excluding environment setup, downloads, and downstream evaluation. Speedups divide comparator and BOLT arithmetic mean runtimes after equal weighting of dataset–budget configurations. Runs use Google Colab with four configured CPU threads. BOLT and Tab-PE main runs use CPU sessions; DP-MERF and DP-HP use T4 GPUs. Timings retain hardware differences without normalization, specified in Appendix C.6.

6 Results

Competitive utility across privacy budgets. BOLT ranks first in one, second in fourteen, and third in one of the sixteen dataset-averaged comparisons of F1 and AUROC across privacy budgets and

Table 1: AUROC by privacy level. Releases are averaged within each dataset–privacy cell (AIM: 3; other methods: 5), then the five datasets are weighted equally, and All ε averages the four budgets. Blue and lilac mark first and second place among synthetic methods using unrounded values; Real is a nonprivate reference excluded from ranking.

Method	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 8$	All ε
Logistic regression					
Real	0.8042	0.8042	0.8042	0.8042	0.8042
AIM	**0.7278**	**0.7518**	**0.7642**	**0.7734**	**0.7543**
BOLT	**0.7298**	**0.7404**	**0.7459**	**0.7484**	**0.7411**
Tab-PE	0.6587	0.6821	0.6917	0.7124	0.6863
DP-MERF	0.6741	0.6722	0.6774	0.6943	0.6795
DP-HP	0.6272	0.6344	0.6455	0.6649	0.6430
LightGBM					
Real	0.8544	0.8544	0.8544	0.8544	0.8544
AIM	**0.7376**	**0.7619**	**0.7687**	**0.7724**	**0.7601**
BOLT	**0.7315**	**0.7493**	**0.7469**	0.7418	**0.7424**
Tab-PE	0.6993	0.7088	0.7302	**0.7494**	0.7219
DP-MERF	0.6909	0.6931	0.6952	0.7101	0.6973
DP-HP	0.5923	0.6506	0.6492	0.6856	0.6444

Table 2: F1 by privacy level, aggregated as in Table 1. Blue and lilac mark first and second place among synthetic methods; Real is excluded from ranking.

Method	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 8$	All ε
Logistic regression					
Real	0.6487	0.6487	0.6487	0.6487	0.6487
AIM	**0.5652**	**0.5858**	**0.5971**	**0.5927**	**0.5852**
BOLT	**0.5552**	**0.5776**	**0.5822**	**0.5842**	**0.5748**
DP-MERF	0.5146	0.5207	0.5326	0.5439	0.5280
Tab-PE	0.4794	0.4985	0.5095	0.5391	0.5066
DP-HP	0.4154	0.4272	0.5109	0.5290	0.4706
LightGBM					
Real	0.6942	0.6942	0.6942	0.6942	0.6942
AIM	**0.5767**	**0.5933**	**0.6067**	**0.6077**	**0.5961**
BOLT	**0.5490**	**0.5797**	**0.5778**	**0.5814**	**0.5720**
DP-MERF	0.4908	0.5205	0.5062	0.5432	0.5152
Tab-PE	0.4799	0.5014	0.5048	0.5273	0.5034
DP-HP	0.3855	0.4003	0.5166	0.5410	0.4608

downstream learners (Tables 1 and 2). Its first-place result is logistic-regression AUROC at $\varepsilon = 0.5$: 0.7298, compared with AIM's 0.7278. Its third-place result is LightGBM AUROC at $\varepsilon = 8$, where Tab-PE scores 0.7494 and BOLT 0.7418. After averaging equally over the twenty dataset-budget configurations, BOLT is second to AIM on all four utility endpoints. Its AUROC deficits are 1.32 and 1.78 percentage points for logistic regression and LightGBM, respectively; its F1 deficits are 1.04 and 2.41 points.

Ranking the five methods separately within each dataset–budget–learner–metric comparison, after averaging releases, gives BOLT a mean utility rank of 2.30 across 80 comparisons, second to AIM's

Table 3: Synthesis runtime relative to BOLT, computed as mean method runtime divided by mean BOLT runtime over the same dataset–privacy cells, so $x\times$ means BOLT is x times faster. Ratios compare the recorded implementations and hardware (Appendix C.6), not hardware-normalized speedups.

Method	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 8$	All ε
BOLT	**1.0**×	**1.0**×	**1.0**×	**1.0**×	**1.0**×
DP-MERF	**13.1**×	**13.2**×	**13.2**×	**12.2**×	**12.9**×
DP-HP	249.7×	252.1×	255.8×	251.9×	252.4×
Tab-PE	314.9×	313.4×	331.7×	324.6×	321.1×
AIM	462.0×	1,230.8×	11,091.5×	13,456.4×	6,516.7×

1.36. These ranks preserve dataset variation: on Shoppers, DP-MERF exceeds BOLT on both LightGBM metrics at $\varepsilon = 0.5, 2, 8$, while BOLT leads at $\varepsilon = 1$. Appendix D reports dataset-level means and release variability. Across the 80 dataset–budget–learner–metric comparisons, BOLT exceeds DP-MERF and Tab-PE in 64 comparisons each, DP-HP in 75, and AIM in 13.

Utility relative to real-data training. For each dataset, budget, learner, and metric, divide the method's mean score by the corresponding Real mean, then average these 80 ratios equally. BOLT retains 87.00% of reference utility, compared with 89.26% for AIM, 79.82% for DP-MERF, 79.00% for Tab-PE, and 72.49% for DP-HP. BOLT's retained utility rises from 84.56% at $\varepsilon = 0.5$ to 87.54%, 87.87%, and 88.03% at $\varepsilon = 1, 2, 8$. Defining the normalized gap as one minus retained utility, BOLT reduces the average gap by 35.58%, 38.09%, and 52.75% relative to DP-MERF, Tab-PE, and DP-HP, respectively. AIM retains the smaller gap. These descriptive summaries use Real as an empirical reference; they are not prediction-error rates. Appendix C.4 specifies the calculations.

Short complete-table synthesis times. BOLT has the lowest mean synthesis time in all twenty core dataset–budget configurations. Its overall mean is 2.85 s, compared with 36.87 s for DP-MERF, the next-fastest method, and 18,587 s (5.16 h) for AIM. Ratios of arithmetic mean runtimes give $12.9\times$ and $6{,}516.7\times$ speed advantages, respectively; the corresponding ratios for DP-HP and Tab-PE are $252.4\times$ and $321.1\times$ (Table 3). Across the 100 BOLT releases, complete tables take 0.53–10.47 s. These comparisons retain the recorded hardware and timing differences described in Appendix C.6.

BOLT combines higher aggregate scores with shorter synthesis times than DP-MERF, DP-HP, and Tab-PE, while AIM provides higher aggregate utility at substantially greater recorded cost. Figure 2 shows the joint comparison.

Conjunctions and classifier labeling. At $\varepsilon = 1$, changing the conjunction count from zero to 512 changes dataset-averaged AUROC/F1 by $+3.87/+4.63$ percentage points for logistic regression and $+8.72/+6.87$ points for LightGBM. Increasing from 512 to 2,048 raises 1 of the four dataset-averaged utility endpoints (Table 9).

Classifier labels minus retained component labels yield mean AUROC/F1 differences of $-0.65/+0.89$ percentage points for logistic regression and $+8.17/+9.91$ for LightGBM. The LightGBM means are driven by datasets where LightGBM performs poorly with retained labels: Adult ($+26.32/+34.50$), Default ($+16.62/+10.48$), and, for F1, Diabetes ($+14.57$) (Tables 9 and 10).

7 Discussion

BOLT demonstrates that a simple covariate generator paired with a richer labeling model can supply useful synthetic training data at low measured synthesis cost. In the evaluated settings, complete tables take seconds to generate while retaining competitive aggregate utility. AIM achieves higher aggregate utility, so BOLT offers a practical tradeoff when shorter synthesis times justify some predictive loss.

The component experiments clarify the choices within this construction. Using 512 conjunctions improves all four dataset-averaged utility endpoints over zero, but increasing to 2,048 improves

only one. These contrasts change both the available interactions and their effective noise. Classifier labeling has small, mixed average effects for logistic regression. Its larger average LightGBM gains concentrate where retained labels perform poorly (Adult and Default for both metrics, and Diabetes for F1), rather than showing uniform improvements. These experiments assess workload and labeling choices within BOLT; they do not isolate its overall advantage over other methods (Tables 9 and 10).

The broader implication concerns where private interaction measurements are spent. AIM and related marginal-based methods use them to reconstruct a joint distribution (McKenna et al., 2019; 2022), and QUAIL privatizes a supervised model separately from its synthesizer (Rosenblatt et al., 2020). BOLT instead derives both models from one jointly calibrated summary and uses the interactions only in the labeling rule, much as model compression uses a teacher to label generated inputs (Buciluǎ et al., 2006). The observed utility and synthesis times indicate that, for a predefined prediction task, interaction measurements can contribute to utility through the labels without a generator that reproduces them.

8 Limitations

BOLT targets binary classification under a fixed public schema and encoding; regression and multi-class tasks remain unevaluated. Its two-component product mixture restricts unconditional covariate dependence, and relabeling cannot restore missing covariate relationships. Numerical means constrain proposal components rather than final classes; infeasible noisy targets are projected. Overlapping Boolean features depart from the naïve-Bayes independence assumption; clipping, shrinkage, and prevalence adjustment do not ensure calibrated real-data probabilities. Synthetic tables need not reproduce the classifier's interaction statistics.

Recovery of the labeling model rests on approximation, coverage, and learning assumptions; linear models on original features need not represent conjunction scores. The population result in Appendix A gives no finite-sample train-synthetic/test-real guarantee. Numerical fields are sampled and evaluated in continuous encoded coordinates; rounding decoded integer or count fields may change predicates and utility, so the reported performance does not extend automatically to rounded tables.

Evidence spans five datasets, four privacy budgets, two learners, and fixed train/test partitions, method configurations, and schema adaptations. F1 and AUROC measure utility for specified labels, without establishing general distributional fidelity; aggregate rankings are descriptive, without establishing statistical non-inferiority. Hardware differs across methods, synthesis times are not hardware-normalized, and timers include checkpoint and progress-reporting overhead differently (Appendix C.6). Baseline settings appear in Appendix C.3.

9 Conclusion

We asked whether private interaction measurements could make synthetic tables useful for a fixed prediction task through their labels, while covariate generation stays simple. In the evaluated settings, they can. Conjunctions enter only BOLT's labeling classifier, yet adding 512 of them improves all four dataset-averaged utility endpoints at $\varepsilon = 1$. Across five datasets and four privacy budgets, BOLT retains 87.0% of real-data reference utility, above DP-MERF, DP-HP, and Tab-PE, and produces a complete table in 2.85 seconds on average, compared with 89.3% and 5.16 hours for AIM. Where this utility difference is acceptable, BOLT reduces the cost of private tabular synthesis from hours to seconds. Multiclass and regression targets, richer covariate proposals, and principled selection of the conjunction workload are natural next steps.

Ethics Statement

Barriers to the adoption of privacy-preserving machine learning are primarily rooted in two challenges: (1) the privacy vs. utility tradeoff, and (2) cost. Privacy-preserving methods are often more expensive to deploy given the additional compute required compared to their non-private counterparts. When applied correctly and in the right situation, methods like BOLT can help to drive the adoption of privacy-preserving methods by reducing the cost relative to existing private synthetic data generation methods. Harms associated with this research relate to practitioner-error (unfaithful implementation yielding untrue privacy guarantees, contextually inappropriate choice of ε or δ resulting in unacceptable levels of privacy leakage), and privacy-washing, where an organization publicly claims to deploy their models with differential privacy to present a privacy-conscious brand image, but chooses privacy budgets that offer little practical protection.

AI use statement

Language models were used to create this work throughout multiple levels of involvement. Core hypotheses, mechanism design, experiments, and measurements were proposed by us. Language models were used to implement these ideas in code. They also provided feedback on methodology and mathematical formulations, assisted with math proofs and statistical analysis, and supported interpretation of results. We were responsible for executing the code and collecting the results, drawing conclusions, and computing the reported descriptive summaries. Language models were employed to draft broad language for the paper, while we read, checked language-to-code alignment, revised wording, statements, conclusions, and generally edited the paper to reflect our own ideas and views. We take responsibility for the final paper.

Reproducibility statement

Sections 3–5 specify the construction, privacy guarantee, and evaluation protocol. Appendix A provides proofs; Appendix B gives numerical details. Appendix C specifies datasets, representations, method settings, randomness, aggregation, and timing procedures. Appendix D reports detailed results and release variability.

References

Miriam Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and Edward Silverman. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26(4): 641–647, 1955. doi: 10.1214/aoms/1177728423.

Borja Balle and Yu-Xiang Wang. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 394–403. PMLR, 2018. URL https://proceedings.mlr.press/v80/balle18a.html.

Matej Balog, Ilya Tolstikhin, and Bernhard Schölkopf. Differentially private database release via kernel mean embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 414–422. PMLR, 2018. URL https://proceedings.mlr.press/v80/balog18a.html.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. URL https://doi.org/10.24432/C5XW20. Dataset.

R. Bock. MAGIC Gamma Telescope. UCI Machine Learning Repository, 2004. URL https://doi.org/10.24432/C52C8B. Dataset.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 535–541. ACM, 2006. doi: 10.1145/1150402.1150464. URL https://www.cs.cornell.edu/~caruana/compression.kdd06.pdf.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography*, volume 9985 of *Lecture Notes in Computer Science*, pp. 635–658. Springer, 2016. doi: 10.1007/978-3-662-53641-4_24.

Dingfan Chen, Raouf Kerkouche, and Mario Fritz. Private set generation with discriminative information. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14678–14690, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/5e1a87dbb7e954b8d9d6c91f6db771eb-Abstract.html.

John Clore, Krzysztof Cios, Jon DeShazo, and Beata Strack. Diabetes 130-US Hospitals for Years 1999–2008. UCI Machine Learning Repository, 2014. URL https://doi.org/10.24432/C5230J. Dataset.

Graham Cormode. Individual privacy vs population privacy: Learning to attack anonymization. arXiv preprint arXiv:1011.2511, 2010. URL https://arxiv.org/abs/1011.2511.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975. doi: 10.1214/aop/1176996454.

Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022. doi: 10.1111/rssb.12454.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss library. arXiv preprint arXiv:2401.08281, 2024. URL https://arxiv.org/abs/2401.08281.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 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*, volume 3876 of *Lecture Notes in Computer Science*, pp. 265–284. Springer, 2006. doi: 10.1007/11681878_14. URL https://people.csail.mit.edu/asmith/PS/sensitivity-tcc-final.pdf.

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11(67):2001–2049, 2010. URL https://www.jmlr.org/papers/v11/ganchev10a.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Frederik Harder, Kamil Adamczewski, and Mijung Park. DP-MERF: Differentially private mean embeddings with random features for practical privacy-preserving data generation. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1819–1827. PMLR, 2021. URL https://proceedings.mlr.press/v130/harder21a.html.

Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu. Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment*, 3(1–2):1021–1032, 2010. doi: 10.14778/1920841.1920970.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456. PMLR, 2015. URL https://proceedings.mlr.press/v37/ioffe15.html.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Ryan McKenna, Daniel Sheldon, and Gerome Miklau. Graphical-model based estimation and inference for differential privacy. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4435–4444. PMLR, 2019. URL https://proceedings.mlr.press/v97/mckenna19a.html.

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. doi: 10.14778/3551793.3551817. URL https://www.vldb.org/pvldb/vol15/p2599-mckenna.pdf.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Michael J. Pazzani. Constructive induction of Cartesian product attributes. In *Information, Statistics and Induction in Science*, pp. 66–77. World Scientific, 1996. URL https://www.ics.uci.edu/~pazzani/Publications/Pazzani-isis96.pdf.

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(85):2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://proceedings.neurips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Lucas Rosenblatt, Xiaoyan Liu, Samira Pouyanfar, Eduardo de Leon, Anuj Desai, and Joshua Allen. Differentially private synthetic data: Applied evaluations and enhancements. arXiv preprint arXiv:2011.05537, 2020. URL https://arxiv.org/abs/2011.05537.

C. Sakar and Yomi Kastro. Online Shoppers Purchasing Intention Dataset. UCI Machine Learning Repository, 2018. URL https://doi.org/10.24432/C5F88Q. Dataset.

Toan Tran, Arturs Backurs, Zinan Lin, Victor Reis, Li Xiong, and Sergey Yekhanin. Differentially private synthetic data via APIs 4: Tabular data. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.08259.

Jaideep Vaidya, Basit Shafiq, Anirban Basu, and Yuan Hong. Differentially private naive Bayes classification. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pp. 571–576. IEEE, 2013. doi: 10.1109/WI-IAT.2013.80. URL https://yhongcs.github.io/pub/wi13.pdf.

Giuseppe Vietri, Cedric Archambeau, Sergul Aydore, William Brown, Michael J. Kearns, Aaron Roth, Ankit Siva, Shuai Tang, and Steven Z. Wu. Private synthetic data for multitask learning and marginal queries. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18282–18295, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/7428310c0f97f1c6bb2ef1be99c1ec2a-Abstract-Conference.html.

Margarita Vinaroz, Mohammad-Amin Charusaie, Frederik Harder, Kamil Adamczewski, and Mi Jung Park. Hermite polynomial features for private data generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22300–22324. PMLR, 2022. URL https://proceedings.mlr.press/v162/vinaroz22a.html.

Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient Monte Carlo. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2493–2502. PMLR, 2015. URL https://proceedings.mlr.press/v37/wangg15.html.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled Rényi differential privacy and analytical moments accountant. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1226–1235. PMLR, 2019. URL https://proceedings.mlr.press/v89/wang19b.html.

Yilin Yang, Kamil Adamczewski, Xiaoxiao Li, Danica J. Sutherland, and Mijung Park. Differentially private neural tangent kernels (DP-NTK) for privacy-preserving data generation. *Journal of Artificial Intelligence Research*, 81:683–700, 2024. doi: 10.1613/jair.1.15985. URL https://www.jair.org/index.php/jair/article/view/15985.

I-Cheng Yeh. Default of Credit Card Clients. UCI Machine Learning Repository, 2009. URL https://doi.org/10.24432/C55S3H. Dataset.

Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. PrivBayes: Private data release via Bayesian networks. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pp. 1423–1434. ACM, 2014. doi: 10.1145/2588555.2588573. URL https://wrap.warwick.ac.uk/id/eprint/63459/.

A PROOFS AND MATHEMATICAL DETAILS

The statements below concern the specified randomized mechanism, conditional on the public schema, feature collection, and other public inputs. They do not require the real-data features to be conditionally independent. That modeling assumption affects the classifier, not the sensitivity or privacy proof.

A.1 EXACT SENSITIVITY OF THE SHARED QUERY

There are d numerical columns and p categorical columns, with $L = d + p > 0$. Numerical coordinates are clipped and encoded in $[-1, 1]$; each categorical column contributes one one-hot block. Consequently,

$$\|g(x)\|_2^2 = \frac{\sum_{k=1}^{d} x_k^2 + p}{L} \le 1.$$

Here and below x_k in this expression denotes the encoded numerical value. Let $z(x) \in \{0,1\}^M$ collect the threshold and conjunction indicators and $a(x) = 2z(x) - \mathbf{1}$. Since $\|a(x)\|_2^2 = M$,

$$\|\psi(x)\|_2^2 = \left\| \left(\frac{g(x)}{\sqrt{2}}, \frac{a(x)}{\sqrt{2M}} \right) \right\|_2^2 = \frac{\|g(x)\|_2^2 + 1}{2} \le 1. \tag{12}$$

The public product domain contains an endpoint record whose numerical coordinates all have absolute value one. At that record equality holds in both bounds, regardless of which Boolean indicators are true.

Proposition 3 (Exact radius and diameter)**.** *For $h(x, y) = 2^{-1/2}(e_y, e_y \otimes \psi(x))$ on the public domain $\mathcal{X} \times \{0, 1\}$,*

$$\sup_{x,y} \|h(x, y)\|_2 = 1, \qquad \sup_{x,y,x',y'} \|h(x, y) - h(x', y')\|_2 = \sqrt{2}.$$

Thus $q(D) = n_0^{-1} \sum_{(x,y)\in D} h(x, y)$ has exact global L_2 sensitivity $1/n_0$ under add/remove adjacency, and $\sqrt{2}/n_0$ under replacement adjacency. The same positive public denominator n_0 is used on both neighboring datasets.

Proof. The squared record norm is $(1 + \|\psi(x)\|_2^2)/2 \le 1$, with equality at an endpoint record. For equal labels, the squared distance is $\|\psi(x) - \psi(x')\|_2^2/2 \le 2$. For different labels the two class blocks are orthogonal, giving squared distance $1 + (\|\psi(x)\|_2^2 + \|\psi(x')\|_2^2)/2 \le 2$. Changing only the label of an endpoint record attains squared distance two. An add/remove edit changes the sum by one record; a replacement changes it by the difference of two records. Division by n_0 gives the stated sensitivities and preserves attainability. □

For fixed-size replacement adjacency, the public sample size can be used as $n_0 = n$. Under add/remove adjacency, neither replacing n_0 by the observed private sample size nor disclosing that size is justified by this proof.

A.2 GAUSSIAN CALIBRATION AND POST-PROCESSING

Theorem 4 (Privacy of the shared release and synthesis)**.** *Fix public n_0, schema, feature collection, output size, and $\varepsilon > 0$, $0 < \delta < 1$. Set $\Delta = 1/n_0$ for add/remove adjacency or $\Delta = \sqrt{2}/n_0$ for fixed-size replacement adjacency. For $\sigma > 0$, if*

$$\Phi\left(\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) - e^{\varepsilon}\Phi\left(-\frac{\Delta}{2\sigma} - \frac{\varepsilon\sigma}{\Delta}\right) \le \delta, \tag{13}$$

then $\widetilde{q} = q(D) + Z$, $Z \sim \mathcal{N}(0, \sigma^2 I)$, is (ε, δ)-differentially private. The joint output consisting of this release, the covariate model, the BOLT classifier, and any fixed finite collection of synthetic tables generated solely from this release also satisfies the same guarantee. The Gaussian randomness is independent of the data and is not disclosed.

Proof. This is the analytic Gaussian mechanism (Balle & Wang, 2018, Theorem 8), specialized to Proposition 3. For completeness, let neighboring query means be u, v with $r = \|u - v\|_2 > 0$. Under $W \sim \mathcal{N}(u, \sigma^2 I)$, the log density ratio is

$$\log \frac{p_u(W)}{p_v(W)} = \frac{\|W - v\|_2^2 - \|W - u\|_2^2}{2\sigma^2} \sim \mathcal{N}\left(\frac{r^2}{2\sigma^2}, \frac{r^2}{\sigma^2}\right).$$

The event attaining $\sup_A \{P_u(A) - e^\varepsilon P_v(A)\}$ is where this log density ratio exceeds ε. Its value is

$$\delta_\varepsilon(r) = \Phi\left(\frac{r}{2\sigma} - \frac{\varepsilon\sigma}{r}\right) - e^\varepsilon \Phi\left(-\frac{r}{2\sigma} - \frac{\varepsilon\sigma}{r}\right).$$

Writing $a = r/(2\sigma) - \varepsilon\sigma/r$ and $b = -r/(2\sigma) - \varepsilon\sigma/r$, the identity $e^\varepsilon \varphi(b) = \varphi(a)$ for the standard normal density φ gives $\delta'_\varepsilon(r) = \varphi(a)/\sigma > 0$. The case $r = 0$ has identical distributions. Since $r \le \Delta$, the condition in (13) bounds the privacy divergence by δ; exchanging u, v gives the other direction.

Given $\widetilde{q}$, all subsequent model construction and sampling use only public inputs and additional randomness independent of both the private data and the Gaussian-release randomness. In particular, projecting each released numerical mean to the feasible interval determined by its repaired bin probabilities and fitting its within-bin exponential tilt require no additional access to D. These operations and subsequent sampling form a Markov kernel applied to $\widetilde{q}$. The post-processing property of differential privacy therefore applies to the entire joint output, including if $\widetilde{q}$ itself is retained (Dwork & Roth, 2014, Proposition 2.1). □

This theorem is a direct application of established privacy results. It covers the stated statistical outputs, not exact private aggregates, private-data evaluation scores, data-dependent execution traces, or arbitrary training telemetry. The mathematical statement uses ideal Gaussian sampling; numerical calibration must satisfy (13) rather than merely terminate a search.

A.3 What the Boolean classifier represents

For repaired probabilities $\widehat{p}_{cj} \in (0, 1)$ and $\widehat{\pi}_c > 0$, the score can be written

$$s(x) = \beta_0 + \sum_{j=1}^{M} \beta_j z_j(x), \tag{14}$$

$$\beta_j = \log \frac{\widehat{p}_{1j}(1 - \widehat{p}_{0j})}{\widehat{p}_{0j}(1 - \widehat{p}_{1j})},$$

$$\beta_0 = \log \frac{\widehat{\pi}_1}{\widehat{\pi}_0} + \sum_{j=1}^{M} \log \frac{1 - \widehat{p}_{1j}}{1 - \widehat{p}_{0j}}.$$

These identities follow by collecting the z_j and constant terms in the Bernoulli log-posterior odds. Each literal is either a base predicate b or $1 - b$. A two-literal conjunction is their product. Hence the logit is a polynomial of degree at most two in the base predicates, restricted to their feasible Boolean patterns. This does not imply that its sigmoid is a degree-two polynomial, that arbitrary high-order interactions are represented, or that the real-data indicators are conditionally independent.

The joint scaling also exposes the cost of a larger feature collection. Before division by noisy class masses, undoing the packing yields class-mass noise with standard deviation $\sqrt{2}\sigma$, raw class-indexed numerical/one-hot sum noise with standard deviation $2\sqrt{L}\sigma$, and signed-indicator sum noise with standard deviation $2\sqrt{M}\sigma$; these sums all use denominator n_0. After division, the conditional estimates are ratios sharing a noisy denominator, so these are not their standard deviations.

A.4 The intercept adjustment

Condition on the shared release and the sampled covariates $X_{1:N}$. Write $s_i = s(X_i)$ and $\rho_* = \mathrm{clip}(\widehat{\pi}_1, 10^{-9}, 1 - 10^{-9})$, matching the numerical target used by the implementation. The sigmoid is $\mathrm{sigmoid}(t) = (1 + e^{-t})^{-1}$. Sampling claims in this and the next subsection describe the ideal randomized sampler. Fixing an implementation sampling seed instead gives a deterministic post-processing map; this does not change the privacy conclusion.

Proposition 5 (Unique adjustment and expected prevalence)**.** *For finite scores and* $0 < \rho_* < 1$*, there is a unique finite* α_N *such that*

$$F_N(\alpha_N) := \frac{1}{N}\sum_{i=1}^{N} \mathrm{sigmoid}(s_i + \alpha_N) = \rho_*.$$

If Y_i *are sampled independently conditional on the release and* $X_{1:N}$ *with probabilities* $r_i = \mathrm{sigmoid}(s_i + \alpha_N)$*, then*

$$\mathbb{E}\left[\frac{1}{N}\sum_i Y_i \,\middle|\, \widetilde{q}, X_{1:N}\right] = \rho_*, \qquad \mathrm{Var}\left(\frac{1}{N}\sum_i Y_i \,\middle|\, \widetilde{q}, X_{1:N}\right) = \frac{1}{N^2}\sum_i r_i(1-r_i) \le \frac{1}{4N}.$$

Proof. F_N is continuous, strictly increasing, and has limits zero and one at $-\infty$ and $+\infty$. The intermediate value theorem gives existence; strict increase gives uniqueness. Conditional expectation and independence of the Bernoulli draws give the two identities. □

The implementation solves this scalar equation numerically, so its mean probability agrees with ρ_* up to numerical error. It agrees with $\widehat{\pi}_1$ when the target clipping is inactive. An exact realized class count is not imposed. The final rows are exchangeable but generally not independent conditional on the release alone: every label probability uses the offset fitted to the whole covariate batch.

Proposition 6 (A constrained KL projection)**.** *Let* $p_i = \mathrm{sigmoid}(s_i)$*. The adjusted probabilities uniquely solve*

$$\min_{r\in[0,1]^N} \sum_{i=1}^{N} D_{\mathrm{KL}}(\mathrm{Ber}(r_i) \| \mathrm{Ber}(p_i)) \quad \textit{subject to} \quad \frac{1}{N}\sum_i r_i = \rho_*. \tag{15}$$

Proof. This is the standard constrained KL-projection construction (Csiszár, 1975; Ganchev et al., 2010), specialized to one mean constraint. The objective is strictly convex, with second derivative $1/[r_i(1-r_i)]$ in coordinate i. Using a multiplier $-\alpha$ on $\sum_i r_i - N\rho_* = 0$, the interior stationarity equations are $\mathrm{logit}(r_i) - \mathrm{logit}(p_i) = \alpha$. Proposition 5 supplies a unique finite α_N whose $r_i = \mathrm{sigmoid}(s_i + \alpha_N)$ are interior and feasible. Convexity and these stationarity equations establish global optimality; strict convexity establishes uniqueness, including against feasible boundary points. □

Thus the adjustment changes Bernoulli probabilities as little as possible in the stated KL direction while matching the expected fraction. It preserves all score orderings. It establishes neither general probability calibration against real labels nor agreement with the released class-conditional interaction statistics.

A.5 Population interpretation and the scope of teacher recovery

Condition on a finite shared release, fixing Q_X, s, and ρ_*. Let

$$F(\alpha) = \mathbb{E}_{Q_X}[\mathrm{sigmoid}(s(X) + \alpha)].$$

Continuity by dominated convergence, strict increase, and the limiting values zero and one give a unique α_* with $F(\alpha_*) = \rho_*$. For iid $X_i \sim Q_X$, the empirical offsets satisfy $\alpha_N \to \alpha_*$ almost surely. To see this, for any $\eta > 0$, $F(\alpha_* - \eta) < \rho_* < F(\alpha_* + \eta)$; the strong law gives the same strict inequalities for F_N eventually. Monotonicity traps α_N between these two endpoints. Applying this argument to positive rational η proves convergence. This argument does not treat the finite-batch labeled rows as iid.

Define a population synthetic law by $X \sim Q_X$ and $Y \mid X \sim \mathrm{Ber}(\eta_*(X))$, where $\eta_*(x) = \mathrm{sigmoid}(s(x) + \alpha_*)$. For any measurable predicted probability $f : \mathcal{X} \to (0,1)$, its population logistic risk satisfies

$$R(f) - R(\eta_*) = \mathbb{E}_{Q_X}[D_{\mathrm{KL}}(\mathrm{Ber}(\eta_*(X)) \| \mathrm{Ber}(f(X)))] \ge 0. \tag{16}$$

The identity follows by subtracting the two conditional Bernoulli cross-entropies and expresses the strict propriety of logarithmic loss (Gneiting & Raftery, 2007). Equality holds precisely when

$f = \eta_* \, Q_X$-almost surely. Equation (14) therefore places the unique probability minimizer in logistic regression on the full Boolean feature vector, including an intercept. Parameters need not be identifiable when features are redundant; the minimizing probabilities are.

For a target covariate distribution P_X, recovery transfers if the target distribution of Boolean feature patterns is absolutely continuous with respect to their distribution under Q_X. Under this condition, a population minimizer in that feature space agrees with $\mathrm{sigmoid}(s + \alpha_*)$ also P_X-almost surely and preserves the teacher's ranking, including ties. Without this support condition, behavior on unseen patterns is unconstrained. This is a population characterization, not a finite-sample guarantee for a fitted learner. It does not apply automatically to logistic regression on raw columns, to a restricted model omitting conjunctions, or to the fixed-penalty optimum of a regularized learner. Nor does it assert that η_* is the real-data posterior.

B Implementation specification

This appendix specifies BOLT's numerical construction. The mechanism operates on binary labels and the public encoded domain $[-1, 1]^d \times \prod_{k=1}^{p} \{1, \ldots, K_k\}$. Set $L = d + p > 0$, and let $M = 7d + J > 0$. The main configuration uses $J = 512$ conjunctions. All constants below are fixed independently of the private observations.

Encoding and Boolean features. Numerical columns are clipped and mapped to $[-1, 1]$ using fixed public transformations. Categorical columns use fixed domains, including an unknown category. Let $v(x)$ concatenate the d encoded numerical values and the categorical one-hot vectors. Then $g(x) = v(x)/\sqrt{L}$. Each numerical column contributes the seven predicates $\mathbf{1}\{x_k > t\}$, with $t \in \{-3/4, -1/2, -1/4, 0, 1/4, 1/2, 3/4\}$. Categorical equality predicates are also available as constituents of conjunctions. They do not contribute standalone terms to the Bernoulli naïve-Bayes score.

The conjunction generator samples two base-predicate indices and two negation bits, rejecting pairs from the same original column. It sorts the two indexed literals and rejects duplicate sorted pairs. Thus uniqueness is syntactic: logically equivalent expressions can remain, for example when one category is the complement of another in a binary column. Sampling is over predicate indices; it is not uniform over column pairs. With the same public seed, the smaller conjunction collections are prefixes of the larger collection. The numerical thresholds followed by these conjunctions form $z(x) \in \{0, 1\}^M$, and $a(x) = 2z(x) - \mathbf{1}$. The active feature map is

$$\psi(x) = \left(\frac{g(x)}{\sqrt{2}}, \frac{a(x)}{\sqrt{2M}} \right).$$

Unpacking the noisy summary. Write the first two coordinates of the joint release as $\widetilde{q}_c^{\,\mathrm{mass}}$ and the subsequent class blocks as $\widetilde{q}_c^{\,\psi}$, for $c \in \{0, 1\}$. The class-mass repair and conditional feature estimates are

$$r_c = \sqrt{2}\, \widetilde{q}_c^{\,\mathrm{mass}}, \qquad m_c = \max\{r_c, 10^{-6}\}, \qquad \widehat{\pi}_c = \frac{m_c}{m_0 + m_1}, \qquad \widehat{\mu}_c = \frac{\sqrt{2}\, \widetilde{q}_c^{\,\psi}}{m_c}.$$

The denominator in $\widehat{\mu}_c$ is the unnormalized repaired mass m_c, not $\widehat{\pi}_c$. Writing $\widehat{\mu}_{c,g}$ and $\widehat{\mu}_{c,a}$ for the two feature blocks, the unscaled signed Boolean expectations are

$$u_{cj} = \mathrm{clip}_{[-1,1]}\big(\sqrt{2M}\, \widehat{\mu}_{c,a,j}\big).$$

These operations remain defined when a class is absent from the input; observing both classes is not a condition of the private-release theorem.

Classifier probabilities. For every Boolean feature, first compute $p_{cj} = \mathrm{clip}_{[10^{-3}, 1-10^{-3}]}((1 + u_{cj})/2)$. Then set $\overline{p}_j = \sum_c \widehat{\pi}_c p_{cj}$ and $\widehat{p}_{cj} = 0.8 p_{cj} + 0.2 \overline{p}_j$. The score uses both $\log(\widehat{p}_{1j}/\widehat{p}_{0j})$ when $z_j(x) = 1$ and $\log((1 - \widehat{p}_{1j})/(1 - \widehat{p}_{0j}))$ when $z_j(x) = 0$, plus the prior log odds $\log(\widehat{\pi}_1/\widehat{\pi}_0)$.

Covariate probabilities. For category r of column k, let $\widehat{\mu}_{c,g,kr}$ denote its one-hot coordinate. The sampler uses

$$w_{c,kr} = \max\{\sqrt{2L}\, \widehat{\mu}_{c,g,kr}, 10^{-4}\}, \qquad Q_{kc}(r) = \frac{w_{c,kr}}{\sum_{r'} w_{c,kr'}}.$$

For numerical column k, collect its seven threshold expectations u_{ckt}. Set $F_{ckt} = 1 - (1 + u_{ckt})/2$ and apply equal-weight isotonic regression to the clipped vector $\mathrm{clip}_{[0,1]}(F_{ck})$. The pool-adjacent-violators algorithm (Ayer et al., 1955) finds

$$\widehat{F}_{ck} = \operatorname*{argmin}_{f_1 \le \cdots \le f_7} \sum_{t=1}^{7} \big(f_t - \mathrm{clip}_{[0,1]}(F_{ckt})\big)^2.$$

Append endpoints $f_0 = 0$ and $f_8 = 1$, take adjacent differences, floor each difference at 10^{-6}, and normalize their sum to one. Denote the resulting eight bin probabilities by ω_{ckb}, $b = 1, \ldots, 8$, on edges $(e_0, \ldots, e_8) = (-1, -3/4, -1/2, -1/4, 0, 1/4, 1/2, 3/4, 1)$. The flooring and renormalization can change the CDF values after isotonic regression. Order-constrained repair of noisy private statistics has precedents in consistency-based inference (Hay et al., 2010); here each numerical marginal is repaired separately, without projection onto jointly realizable feature moments. The sampler does not use the classifier's pooled shrinkage.

Numerical values within bins. Undo the first-moment scaling to obtain $\widehat{v}_{ck} = \sqrt{2L}\,\widehat{\mu}_{c,g,k}$. With the bin probabilities fixed, project this target onto the closure of the attainable mean range:

$$\underline{v}_{ck} = \sum_b \omega_{ckb} e_{b-1}, \qquad \overline{v}_{ck} = \sum_b \omega_{ckb} e_b, \qquad v^*_{ck} = \mathrm{clip}_{[\underline{v}_{ck}, \overline{v}_{ck}]}(\widehat{v}_{ck}).$$

For an interior target, choose one scalar τ_{ck} shared across bins and use

$$Q_{kc}(dx \mid B = b) = \frac{\exp(\tau_{ck} x)}{\int_{e_{b-1}}^{e_b} \exp(\tau_{ck} t)\, dt}\, dx, \qquad e_{b-1} < x < e_b, \qquad \Pr(B = b \mid C = c) = \omega_{ckb}.$$

Writing $w_b = e_b - e_{b-1}$ and $A(z) = (1 - e^{-z})^{-1} - z^{-1}$ for $z \neq 0$, with $A(0) = 1/2$, the scalar equation is

$$\sum_b \omega_{ckb}\{e_{b-1} + w_b A(\tau_{ck} w_b)\} = v^*_{ck}.$$

Its left side is strictly increasing, with limits $\underline{v}_{ck}$ and $\overline{v}_{ck}$. Zero tilt gives a uniform conditional distribution within every bin. Starting from $[0, 1]$ or $[-1, 0]$ according to the target's position relative to the zero-tilt mean, the implementation doubles the relevant bracket endpoint, allowing at most 128 checks per endpoint, then performs 80 bisection steps. It samples by inverse CDF using one uniform draw per numerical value. Taylor expansions for $|\tau_{ck} w_b| < 10^{-3}$ in the mean and $|\tau_{ck} w_b| < 10^{-4}$ in the inverse CDF, together with reflected exponential evaluations elsewhere, avoid cancellation and overflow.

A target at either feasible endpoint uses the corresponding limit $\tau_{ck} \to -\infty$ or $+\infty$. Floating-point values are moved strictly inside their selected bins, including after inverse and forward transformation, so threshold predicates retain the sampled bin membership. The initial inset uses the nearest interior floating-point values; if the transformation loses that inset, the repair starts at four machine epsilons in encoded units and doubles it, capped at one quarter of the bin width, for at most 16 checks. Thus mean matching describes the distributional expectation, up to numerical precision, rather than an exact finite-sample constraint.

Sample a latent component $C_i \sim \mathrm{Categorical}(\widehat{\pi})$ and sample the original columns independently given C_i, producing $Q_X = \sum_c \widehat{\pi}_c \prod_k Q_{kc}$. The numerical mean constraints apply conditional on C_i; they need not hold conditional on the final labels Y_i. Conjunction statistics affect the classifier, not the covariate proposal. All fitting and sampling use only the shared noisy release and public quantities.

Final labels. With temperature one and scores $s_i = s(X_i)$, set the prevalence target to $\rho_* = \mathrm{clip}_{[10^{-9}, 1-10^{-9}]}(\widehat{\pi}_1)$ and solve $N^{-1} \sum_i \mathrm{sigmoid}(s_i + \alpha) = \rho_*$. Use 120 bisection steps in $[-\max_i s_i - 60, -\min_i s_i + 60]$. Then sample the final labels independently conditional on the complete covariate batch and summary, with success probabilities $\mathrm{sigmoid}(s_i + \alpha)$. The component labels C_i are discarded. No exact final class count is enforced. In the paired relabeling ablation, the same covariates and private summary are also evaluated with $Y_i = C_i$.

Public parameters and release scope. The experiment sets $N = n_0$, where n_0 is a fixed public benchmark reference size. Under add/remove adjacency this constant is unchanged on neighboring inputs; it is not estimated from the input length. The main grid uses $\varepsilon \in \{0.5, 1, 2, 8\}$ and $\delta = n_0^{-2}$. The rule collection uses the fixed public seed $20260813 + 424242$. Synthetic sampling uses a public replicate seed. Privacy noise is generated from separate unpublished randomness.

The Gaussian-scale routine numerically solves the analytic Gaussian calibration equation with 300 bisection steps on $[10^{-3}\Delta, 10^{4}\Delta]$. That bracket covers the configured study, and is fixed rather than adaptive. The theorem describes the real-valued Gaussian mechanism with a scale satisfying its calibration inequality; numerical evaluation uses floating-point aggregation and pseudorandom sampling.

Encoded output and downstream evaluation. The active schema marks every encoded numerical column as continuous. Consequently, the synthetic output is evaluated before rounding count or integer fields into a raw-table schema. Utility after rounding has not been established by the encoded-output experiment. The ordinary downstream classifiers receive encoded numerical values and categorical one-hot vectors, without the threshold and conjunction feature expansion. A population result about logistic learning in the teacher's Boolean feature space therefore does not automatically apply to that ordinary downstream logistic-regression endpoint.

C Reproducibility

C.1 Datasets and fixed representation

We retain the training and test partitions of the archived public UCI dataset snapshot (snapshot identifier withheld for anonymous review). Adult retains the official training/test partition. Default, MAGIC, and Shoppers retain the archived 90/10 partitions generated with seed 42. Diabetes retains the archived 60% training and 20% test partitions of the 101,766-record source; the separate 20% validation partition is not added to training. We preserve the archived row order and do not reconstruct these partitions from current versions of the original datasets.

Table 4: Core classification datasets. Numerical and categorical counts refer to original predictors. Encoded width includes one indicator for every category, including the unknown category; the response is excluded.

Dataset	Training	Test	Numerical	Categorical	Encoded width
Adult (Becker & Kohavi, 1996)	32,561	16,281	5	9	129
Default (Yeh, 2009)	27,000	3,000	14	9	102
Diabetes (Clore et al., 2014)	61,059	20,354	9	27	2,347
MAGIC (Bock, 2004)	17,117	1,902	10	0	10
Shoppers (Sakar & Kastro, 2018)	11,097	1,233	7	10	132

The positive responses are income above \$50,000 for Adult; default payment next month for Default; any readmission (within or after 30 days) versus no readmission for Diabetes; gamma signal (`g`) versus hadronic background (`h`) for MAGIC; and a purchasing session (`Revenue=true`) for Shoppers. Labels are mapped to zero and one. Label text is stripped of surrounding whitespace, compared without case distinctions, and stripped of a final period when present. Boolean and numerical spellings of zero and one are accepted for Shoppers; numerical zero/one spellings are accepted for Default and Diabetes.

Numerical representation. Each numerical field with interval $[a, b]$ is represented as $2(\mathrm{clip}_{[a,b]}(x) - a)/(b - a) - 1$. Intervals are fixed across methods, privacy budgets, and replicates. No empirical extrema, quantiles, moments, or imputation values are fitted. Nonfinite or nonnumeric entries in numerical fields are rejected. The following lists also specify the order of the numerical coordinates:

- Adult: `age` $[17, 90]$; `fnlwgt` $[0, 1{,}500{,}000]$; `capital.gain` $[0, 100{,}000]$; `capital.loss` $[0, 5{,}000]$; `hours.per.week` $[1, 99]$.

- Default: LIMIT_BAL $[0, 2{,}000{,}000]$; AGE $[18, 100]$; BILL_AMT1–BILL_AMT6, each $[-1{,}000{,}000, 2{,}000{,}000]$; PAY_AMT1–PAY_AMT6, each $[0, 2{,}000{,}000]$.
- Diabetes: num_lab_procedures $[0, 200]$; num_medications, number_outpatient, number_emergency, and number_inpatient, each $[0, 100]$; age $[0, 9]$; time_in_hospital $[1, 14]$; num_procedures $[0, 10]$; number_diagnoses $[0, 20]$.
- MAGIC: Length and Width, each $[0, 500]$; Size $[0, 10]$; Conc and Conc1, each $[0, 1]$; Asym $[-1{,}000, 1{,}000]$; M3Long and M3Trans, each $[-500, 500]$; Alpha $[0, 100]$; Dist $[0, 1{,}000]$.
- Shoppers: Administrative_Duration $[0, 5{,}000]$; Informational_Duration $[0, 3{,}000]$; ProductRelated $[0, 1{,}000]$; ProductRelated_Duration $[0, 100{,}000]$; BounceRates and ExitRates, each $[0, 1]$; PageValues $[0, 1{,}000]$.

BOLT numerical outputs are evaluated in this continuous representation, including fields that originally represented integer counts. We do not round them back to raw integer values before evaluation.

Categorical representation. Surrounding whitespace is removed from category strings. Literal strings such as NA, nan, ?, and empty are retained as strings, rather than interpreted as numerical missing values. Every field has an additional final category __UNKNOWN__; strings outside its fixed domain map to this category. All categories receive indicator coordinates, without dropping a reference category.

For Diabetes and Shoppers, each categorical domain is the lexicographically sorted union of its distinct strings in the archived public training and test partitions, followed by the unknown category. Thus this benchmark treats the fixed public snapshot, including its category vocabulary, as public side information; domain discovery on confidential records is outside the privacy claim. The categorical fields, in coordinate order, are:

- Diabetes: race, gender, admission_type_id, discharge_disposition_id, admission_source_id, diag_1, diag_2, diag_3, max_glu_serum, A1Cresult, metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, tolazamide, insulin, glyburide-metformin, change, diabetesMed.
- Shoppers: Administrative, Informational, SpecialDay, Month, OperatingSystems, Browser, Region, TrafficType, VisitorType, Weekend.

Adult uses the following ordered fixed domains, each followed by the unknown category. Field order is the order of the list:

- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- education.num: integers 1 through 16 in increasing order.
- marital.status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: Female, Male.

- `native.country`: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

The original category spellings above are retained. Default's categorical fields, in order, are `SEX`, `EDUCATION`, `MARRIAGE`, `PAY_0`, `PAY_2`, `PAY_3`, `PAY_4`, `PAY_5`, and `PAY_6`. Their respective fixed domains are integers $1, 2$; $0, \ldots, 6$; $0, \ldots, 3$; and $-2, \ldots, 8$ for each payment-status field, all in increasing order, followed by the unknown category. MAGIC has no categorical predictors.

C.2 Privacy and replication

Each core cell is defined by a dataset, synthesis method, and privacy budget. The public reference size n_0 is the training-set size in Table 4; both the normalization constant and the synthetic output size $N = n_0$ remain fixed under add/remove adjacency. The primary experiments use $\varepsilon \in \{0.5, 1, 2, 8\}$ and $\delta = n_0^{-2}$. There are five independent releases per cell for BOLT, DP-MERF, DP-HP, and Tab-PE, and three for AIM. Replicate indices are $0, \ldots, 4$ and $0, \ldots, 2$, respectively. Five nonprivate Real evaluations per core dataset provide 25 reference observations. The 460 private outputs are complete.

Public seeds fix the rule collection, evaluation partitions, and, where used, model initialization and synthetic sampling. Privacy noise is never seeded from a disclosed seed. AIM and Tab-PE use unpublished randomness throughout private synthesis. Replicates are independent private measurements, each at the stated privacy budget; composition across releases follows Section 4.

C.3 Synthesis and downstream evaluation settings

C.3.1 BOLT configuration

BOLT uses the construction and numerical constants in Section 3 and Appendix B: seven thresholds per numerical predictor, $J = 512$ conjunctions, shrinkage 0.2, and temperature one. The conjunction collection uses public seed 20685055; the same seed defines the nested collections for component experiments. Encoded numerical outputs are evaluated without rounding to raw integer units. All 100 main-grid BOLT outputs use the released numerical first moments to shape values within the threshold-defined bins. For each proposal class and numerical predictor, the released mean is projected onto the feasible interval determined by the repaired bin probabilities. A common exponential tilt across that column's bins matches the projected mean in expectation while keeping the bin probabilities fixed; feasible endpoints use the corresponding limiting distributions. This mean constraint concerns proposal components, not exact finite-sample means or means conditional on final labels. The fitting and sampling use only the shared private summary and incur no additional privacy expenditure.

C.3.2 Comparator configurations

The comparisons use a shared public representation, privacy convention, output size, and evaluator. They assess the following fixed configurations, rather than independently optimized configurations for every dataset.

AIM. We adapt AIM (McKenna et al., 2022) with Private-PGM (McKenna et al., 2019) to the shared representation. Numerical variables are discretized into 16 equal-width bins on $[-1, 1]$. The workload comprises singleton and pairwise marginals, including those involving the label, with at most 10,000 cells per marginal. Private-PGM uses a maximum model size of 80 MiB, 1,000 iterations for adaptive fits, and 2,500 for the final fit. The adaptive AIM accountant converts the target (ε, δ) to a zero-concentrated privacy budget ρ (Bun & Steinke, 2016) by numerically inverting the optimized conversion bound in its included accountant. Specifically, it minimizes $\exp\{(\alpha - 1)(\alpha\rho - \varepsilon)\}(1 - 1/\alpha)^\alpha/(\alpha - 1)$ over $\alpha \geq 1.01$ and chooses the largest ρ for which the resulting bound is at most δ, using 1,000 bisection iterations for each search. Its nominal initial round count is 16 times the number of attributes, including the label; Gaussian measurement and private query selection receive 90% and 10% of each adaptive round's ρ allocation, respectively;

initial singleton measurements spend only Gaussian-query budget. The adaptive schedule increases measurement accuracy when changes in the estimated marginals fall below its noise criterion. Sampled numerical bins are decoded to their midpoints. The core optimizer budgets are not reduced to meet a runtime target.

Common neural-generator settings. DP-MERF and DP-HP map encoded numerical values to $[0, 1]$ and concatenate categorical one-hot vectors. Let D denote this full input dimension. Their conditional generator maps ten independent standard-normal latent coordinates and a binary class label through layers of width $11 \to 4D \to 2D \to D$, with batch normalization (Ioffe & Szegedy, 2015) and ReLU after each hidden layer. Numerical outputs use sigmoid heads; each categorical column has a separate softmax head and is decoded by argmax. Numerical outputs are mapped back to $[-1, 1]$.

Both methods use Adam (Kingma & Ba, 2015) with learning rate 0.01, moments $(0.9, 0.999)$, stabilizer 10^{-8}, and no weight decay. The batch size is $\max\{2, \mathrm{round}(0.1n_0)\}$: 3,256, 2,700, 6,106, 1,712, and 1,110 for Adult, Default, Diabetes, MAGIC, and Shoppers, respectively. Private class probabilities are floored at 0.001 and renormalized. Each optimization batch draws all but two labels from these probabilities and appends one label of each class. Final synthesis draws all labels from the private probabilities without forcing both classes to occur. Feature computations are accumulated in chunks of 256 while retaining complete generator batches for batch normalization. Bandwidth is fixed at one; no bandwidth is estimated from confidential examples.

DP-MERF. Our schema-adapted DP-MERF configuration follows the private random-feature matching approach of Harder et al. (2021). The numerical block uses 2,000 random Fourier features (Rahimi & Recht, 2007) from 1,000 Gaussian frequencies. When categorical predictors are present, their one-hot block, divided by the square root of its dimension, is appended. Two Gaussian queries release class proportions and class-indexed feature means, jointly accounted for at the target privacy budget. A common Gaussian noise multiplier is calibrated for their composition with AutoDP (Wang et al., 2019); each query's standard deviation is that multiplier times its sensitivity. For add/remove adjacency, their sensitivities are $1/n_0$ and B/n_0, where $B = \sqrt{2}$ when both blocks are present and $B = 1$ otherwise. Feature means are divided by the released class probabilities. The generator minimizes squared error between private and generated class-conditional feature estimates for 200 updates, with a constant learning rate. The generated class-block sums use batch size times the corresponding released class probability as denominator.

DP-HP. Our schema-adapted DP-HP configuration follows Vinaroz et al. (2022). The sum kernel uses Hermite order 100, giving 101 basis terms per encoded coordinate, normalized by $\sqrt{D}$ and projected to the unit Euclidean ball. Each of 20 epochs selects five distinct encoded coordinates and constructs product features of order five, giving $6^5 = 7,776$ terms, also projected to the unit ball. Each epoch has ten optimizer updates. The product loss weight is 0.1, and the learning rate is multiplied by 0.9 after each epoch. The count and sum-feature queries together receive 0.6ε and $\delta/2$; the 20 product queries together receive 0.4ε and $\delta/2$. Gaussian composition is calibrated within each group, followed by additive composition of the two groups, giving 22 queries in total. Product statistics are aggregated separately by class. Generated sum-feature estimates divide by the realized synthetic class counts, whereas generated product estimates divide by batch size times the released class probabilities.

Tab-PE. We adapt the tabular Private Evolution method of Tran et al. (2026) to the shared representation. Class-conditional populations undergo 29 private nearest-neighbor voting updates. Private class counts receive 0.02ε and 0.02δ; the remaining budget accounts for the 29 Gaussian-noised voting histograms. For add/remove adjacency, a voting histogram has sensitivity one. Using Gaussian-DP composition (Dong et al., 2022), if μ is the parameter corresponding to $(0.98\varepsilon, 0.98\delta)$, each of the 29 voting histograms uses noise standard deviation $\sqrt{29}/\mu$. The private class allocation sums to $N = n_0$, with at least one initial candidate per class.

The distance embedding uses numerical values mapped to $[0, 1]$ with weight one and categorical one-hot entries with weight $1/3$. Exact squared Euclidean searches use CPU FAISS (Douze et al., 2024). The first four updates sample parents with replacement in proportion to nonnegative noisy votes, with uniform sampling if their sum is zero, then replace each selected parent with one mutated

copy. Subsequent updates rank the raw noisy votes; in this phase, each selected parent contributes itself and three independently mutated copies to the next candidate population. The final table contains the last selected parents, shuffled across classes.

At private update $t \in \{1, \ldots, 29\}$, the mutation rate is

$$r_t = 0.5 - 0.49 \left(\frac{t-1}{30} \right)^{0.2}.$$

Categorical values are replaced uniformly within their public domains with probability r_t. Numerical coordinates are initialized uniformly on the continuous interval $[-1, 1]$ and mutated by adding a uniform draw on $[-2r_t, 2r_t]$, then clipping to $[-1, 1]$ without integer rounding. All 100 main-grid Tab-PE outputs use this continuous numerical representation, including coordinates whose original fields were integer counts. The final executed rate is approximately 0.016715; 0.01 is the unexecuted endpoint of the schedule.

C.3.3 Downstream prediction and evaluation randomness

Both downstream learners receive the numerical coordinates followed by the categorical indicator coordinates specified above. BOLT's Boolean predicates and conjunctions are not supplied to these learners. We use neither additional feature standardization nor class or sample weights, hyperparameter search, or holdout-based early stopping.

The scikit-learn implementation (Pedregosa et al., 2011) of logistic regression uses an intercept, an ℓ_2 penalty with inverse regularization strength $C = 1$, L-BFGS optimization, tolerance 10^{-4}, and at most 2,000 iterations. LightGBM uses a binary logistic objective, traditional gradient-boosted trees, 100 boosting rounds, learning rate 0.1, at most 31 leaves per tree, and no explicit depth limit. Its minimum leaf count is 20, minimum leaf Hessian sum is 10^{-3}, and minimum split gain is zero. It uses all rows and columns, with no subsampling, and zero ℓ_1 and ℓ_2 penalties on leaf values. Histogram construction uses at most 200,000 rows. The maximum bin count is 255, with a minimum of three observations per bin. LightGBM uses deterministic row-wise computation, four CPU threads, and fixed seed 20260813. These settings correspond to scikit-learn 1.8.0 and LightGBM 4.7.0.

For each synthetic table, we process class zero followed by class one, shuffle each class's indices from their original row order, and hold out $h_c = \min\{n_c - 1, \max\{1, \mathrm{round}(0.2n_c)\}\}$ examples from class c, where rounding uses ties to even. The remaining examples form the classifier-training partition. The two class-specific training subsets are concatenated and shuffled; the two holdout subsets are concatenated and shuffled separately. Both learners use these same partitions. The dense synthetic table is evaluated directly, without resampling rows. Tables with fewer than two examples of either class yield an unavailable utility endpoint; we neither change their labels nor redraw the private release.

For each fitted learner, the F1 threshold maximizes positive-class F1 on its synthetic holdout over $0.02, 0.03, \ldots, 0.98$. Ties select the smallest threshold. An example is predicted positive when its predicted probability is at least the selected threshold. We evaluate this F1 and probability-based AUROC on the archived real test partition. The Real reference instead shuffles the real training indices without stratification, fits on the first $\lfloor 0.8n \rfloor$ indices, and uses the remainder for threshold selection. Its final test partition is identical to that used for synthetic-data evaluation. Test responses are never used for threshold or hyperparameter selection.

Replicate indices are $0, \ldots, 4$; AIM uses $0, 1, 2$. The evaluation permutation seed is shared across methods and privacy budgets for a given dataset and replicate. To derive it, join a fixed benchmark tag (withheld for anonymous review), the string `EVALUATION`, the lower-case dataset name, `classification`, the decimal input and output row counts, and the decimal replicate index with ASCII unit separators (U+001F). Take the first four bytes of the SHA-256 digest of this UTF-8 string as a little-endian unsigned integer and initialize a NumPy PCG64 generator with it. The same rule supplies the Real permutation seed. Independent private releases use separate unpublished randomness; these public evaluation seeds do not seed privacy noise.

C.4 AGGREGATION AND REPORTING

Raw-score tables analyze learners and metrics separately. For method m, dataset d, privacy budget e, learner ℓ, and metric k, let $S_{mde\ell kr}$ denote the real-test score of replicate r. First compute the arithmetic replicate mean $\overline{S}_{mde\ell k} = R_m^{-1} \sum_r S_{mde\ell kr}$. The per-budget aggregate is $5^{-1} \sum_d \overline{S}_{mde\ell k}$; the across-budget aggregate averages the four budget-specific values equally. Consequently, larger datasets and methods with more replicates do not receive greater weight. Headline table rankings use unrounded dataset-averaged scores and dense ranks; the sixteen comparisons comprise four privacy budgets, two learners, and two metrics. The separate mean utility rank instead ranks the five private methods within each of the 80 dataset–budget–learner–metric comparisons, then averages their ranks equally. Higher utility receives a lower rank, ties receive their average occupied rank, and Real is excluded. Reported cell standard deviations are sample standard deviations across independent releases, not confidence intervals.

For reference-normalized utility, let $\overline{S}_{\mathrm{Real},d\ell k}$ be the corresponding mean of the five Real reference replicates, reused across privacy budgets. Define

$$Q_m = \frac{1}{80} \sum_{d,e,\ell,k} \frac{\overline{S}_{mde\ell k}}{\overline{S}_{\mathrm{Real},d\ell k}}, \qquad G_m = 1 - Q_m.$$

Thus $100Q_m$ is the percentage of reference utility retained and $100G_m$ is the normalized utility gap. Per-budget summaries average the corresponding 20 ratios. Normalization precedes averaging across datasets and endpoints. BOLT's percentage gap reduction relative to comparator c is $100(1 - G_{\mathrm{BOLT}}/G_c)$, a reduction in the averaged normalized gap rather than an average of cell-wise reductions. These are descriptive summaries with equal weights for F1 and AUROC and for both learners. Real is an empirical reference, not a theoretical ceiling, and the gap is not a prediction-error probability; in particular, AUROC's chance reference is 0.5, not zero. Ratios are not clipped, although no evaluated private-method cell mean exceeds its Real reference.

C.5 COMPONENT EXPERIMENTS

The component experiment contains 125 outputs from 100 private release jobs, including 25 matched labeling pairs. At $\varepsilon = 1$, five releases per dataset use each of the nested conjunction collections of sizes 0, 128, 512, and 2,048. All configurations use the numerical sampler in Appendix B, with released first moments determining the within-bin distributions. The paired 512-conjunction control shares its private summary and generated covariates with the retained-component-label output. Labeling effects are computed within each pair before dataset averaging. The paired classifier-label control supplies the component reference separately from the main-grid results. Tables 9 and 10 report the corresponding utility and paired differences.

C.6 RUNTIME MEASUREMENT AND COMPUTING ENVIRONMENT

The synthesis endpoint is elapsed computation for private measurement/model fitting plus production of the first complete N-row synthetic table. It includes BOLT's whole-batch prevalence adjustment. It excludes environment installation, data downloads, downstream learner fitting, final evaluation, and report generation. Input loading and preparation of the public predicate collection precede the synthesis timers. It is neither sample-only throughput nor latency to an independently requested single row. Separately recorded private-query times for neural methods are not added again to fitting time.

For cell runtime T_{mder}, compute $\overline{T}_{mde} = R_m^{-1} \sum_r T_{mder}$ and $\overline{T}_m = 20^{-1} \sum_{d,e} \overline{T}_{mde}$. The reported overall speedup against comparator m is $\overline{T}_m / \overline{T}_{\mathrm{BOLT}}$. Per-budget runtime summaries average over the five datasets instead. These are ratios of arithmetic means, not averages or geometric means of cell-wise speed ratios. Speed–utility plots use the corresponding arithmetic mean latency and display F1 and AUROC separately.

The main-grid BOLT and Tab-PE results were obtained in high-RAM CPU-only Google Colab sessions. All 100 outputs for each method record an Intel Xeon CPU at 2.20 GHz and four configured CPU threads. DP-MERF and DP-HP were measured on Intel Xeon CPUs at 2.00 GHz with NVIDIA Tesla T4 GPUs (15,360 MiB). AIM has 34 measurements from Intel Xeon CPUs at 2.00 GHz, one

from an Intel Xeon CPU at 2.20 GHz, and 20 from AMD EPYC 7B12 CPU sessions. Hardware metadata are unavailable for five AIM outputs and five Real reference observations. All 100 BOLT component jobs were measured on Intel Xeon CPUs at 2.20 GHz. Four CPU threads were configured per job. BOLT, AIM with a NumPy backend, Tab-PE with exact CPU FAISS search, and both downstream learners execute on CPU; DP-MERF and DP-HP execute through CUDA. Reported runtimes retain these observed environments without hardware normalization. For resumed fits, completion hardware need not identify every earlier execution session.

Restart bookkeeping introduces additional timing qualifications. Tab-PE excludes the overhead of reporting progress. Neural fit timers exclude checkpoint-write overhead where tracked. AIM records active fitting and associated recorded checkpoint/progress work, excludes offline downtime, and can omit work lost after the most recent saved state. Complete-table times retain these measurement boundaries; they are not corrected into a hypothetical uniform timing protocol.

Where software metadata are recorded, the shared base versions are Python 3.13.15, NumPy 2.1.3, SciPy 1.16.3, pandas 2.2.3, scikit-learn 1.8.0, LightGBM 4.7.0, PyTorch (Paszke et al., 2019) 2.11.0, AutoDP 0.2.3.1, FAISS CPU 1.15.0, and Private-PGM 0.0.1. Recorded PyTorch builds include both CPU-only and CUDA 12.8 builds. The BOLT and Tab-PE main-grid sessions have the CUDA-enabled build installed but execute on CPU; the installed build does not indicate GPU use. Complete software metadata are unavailable for the five AIM outputs and five Real observations noted above. Archived observations are included only when the dataset, public representation, method configuration, privacy parameters, output size, and replicate match the final experimental specification; test utility is not a selection criterion.

D Additional Results

Sample standard deviations describe variation across releases or paired differences. CI's were not computed.

D.1 Utility by dataset and privacy budget

The core grid contains 460 completed private outputs and 25 Real reference observations. BOLT exceeds comparator cell means over dataset, privacy budget, learner, and metric combinations as follows: AIM: 13/80; DP-MERF: 64/80; DP-HP: 75/80; Tab-PE: 64/80. These are descriptive counts, not independent significance tests. All 100 BOLT main-grid outputs use the released numerical first moments to determine the within-bin sampling distributions, followed by Boolean-teacher labeling.

Table 5: Dataset-level predictive utility at $\varepsilon = 0.5$. Entries are arithmetic means (sample standard deviations) across independent releases; observations per cell: Real: 5; BOLT: 5; AIM: 3; DP-MERF: 5; DP-HP: 5; Tab-PE: 5. Real is repeated across privacy-budget tables. LR denotes logistic regression. F1 is positive-class F1; standard deviations are descriptive, not confidence intervals.

Method	LR AUROC	LR F1	LightGBM AUROC	LightGBM F1
Adult				
Real	0.9048 (0.000)	0.6831 (0.001)	0.9265 (0.000)	0.7193 (0.002)
BOLT	0.8691 (0.003)	0.6326 (0.005)	0.8641 (0.007)	0.6175 (0.015)
AIM	0.8782 (0.005)	0.6270 (0.022)	0.8852 (0.003)	0.6635 (0.005)
DP-MERF	0.8163 (0.020)	0.4906 (0.060)	0.8036 (0.021)	0.3742 (0.086)
DP-HP	0.7148 (0.068)	0.4210 (0.096)	0.6647 (0.086)	0.3715 (0.064)
Tab-PE	0.8160 (0.018)	0.5405 (0.034)	0.8377 (0.010)	0.5758 (0.007)
Default				
Real	0.7480 (0.001)	0.5117 (0.004)	0.7744 (0.002)	0.5256 (0.006)
BOLT	0.7165 (0.010)	0.4648 (0.009)	0.7166 (0.008)	0.4658 (0.017)
AIM	0.7152 (0.019)	0.4840 (0.021)	0.7199 (0.012)	0.4903 (0.014)
DP-MERF	0.6941 (0.014)	0.4885 (0.010)	0.6965 (0.011)	0.4634 (0.018)
DP-HP	0.5520 (0.115)	0.1676 (0.181)	0.5389 (0.141)	0.1649 (0.188)
Tab-PE	0.6516 (0.069)	0.3452 (0.140)	0.6729 (0.028)	0.3971 (0.014)
Diabetes				
Real	0.6824 (0.001)	0.6536 (0.001)	0.7044 (0.001)	0.6622 (0.001)
BOLT	0.5667 (0.012)	0.6041 (0.010)	0.5691 (0.010)	0.6038 (0.005)
AIM	0.5821 (0.011)	0.6305 (0.007)	0.5907 (0.013)	0.6342 (0.002)
DP-MERF	0.5024 (0.015)	0.6294 (0.003)	0.5082 (0.009)	0.5166 (0.251)
DP-HP	0.5063 (0.030)	0.5570 (0.129)	0.5141 (0.015)	0.4759 (0.261)
Tab-PE	0.4969 (0.015)	0.5188 (0.050)	0.5213 (0.017)	0.4449 (0.165)
MAGIC				
Real	0.8450 (0.000)	0.8461 (0.002)	0.9484 (0.001)	0.9185 (0.003)
BOLT	0.8171 (0.003)	0.8206 (0.003)	0.8157 (0.004)	0.8202 (0.003)
AIM	0.8339 (0.005)	0.8293 (0.005)	0.8558 (0.010)	0.8437 (0.009)
DP-MERF	0.7617 (0.067)	0.8169 (0.031)	0.7438 (0.060)	0.8014 (0.026)
DP-HP	0.7483 (0.014)	0.7912 (0.012)	0.7445 (0.028)	0.8055 (0.014)
Tab-PE	0.7594 (0.021)	0.7839 (0.000)	0.8333 (0.015)	0.8369 (0.017)
Shoppers				
Real	0.8408 (0.003)	0.5489 (0.012)	0.9181 (0.002)	0.6452 (0.008)
BOLT	0.6795 (0.018)	0.2536 (0.035)	0.6921 (0.031)	0.2379 (0.069)
AIM	0.6293 (0.021)	0.2554 (0.009)	0.6362 (0.033)	0.2517 (0.003)
DP-MERF	0.5959 (0.056)	0.1478 (0.049)	0.7026 (0.109)	0.2985 (0.100)
DP-HP	0.6143 (0.053)	0.1404 (0.069)	0.4994 (0.239)	0.1095 (0.074)
Tab-PE	0.5699 (0.061)	0.2087 (0.077)	0.6311 (0.073)	0.1450 (0.068)

Table 6: Dataset-level predictive utility at $\varepsilon = 1$. Entries are arithmetic means (sample standard deviations) across independent releases; observations per cell: Real: 5; BOLT: 5; AIM: 3; DP-MERF: 5; DP-HP: 5; Tab-PE: 5. Real is repeated across privacy-budget tables. LR denotes logistic regression. F1 is positive-class F1; standard deviations are descriptive, not confidence intervals.

Method	LR AUROC	LR F1	LightGBM AUROC	LightGBM F1
Adult				
Real	0.9048 (0.000)	0.6831 (0.001)	0.9265 (0.000)	0.7193 (0.002)
BOLT	0.8709 (0.001)	0.6376 (0.003)	0.8662 (0.003)	0.6224 (0.006)
AIM	0.8816 (0.004)	0.6386 (0.006)	0.8920 (0.005)	0.6742 (0.004)
DP-MERF	0.8154 (0.015)	0.4781 (0.044)	0.8021 (0.028)	0.3625 (0.060)
DP-HP	0.7265 (0.043)	0.4889 (0.029)	0.6725 (0.050)	0.3930 (0.030)
Tab-PE	0.8340 (0.007)	0.5702 (0.011)	0.8568 (0.005)	0.6089 (0.010)
Default				
Real	0.7480 (0.001)	0.5117 (0.004)	0.7744 (0.002)	0.5256 (0.006)
BOLT	0.7232 (0.004)	0.4593 (0.010)	0.7117 (0.026)	0.4478 (0.010)
AIM	0.7349 (0.004)	0.5032 (0.015)	0.7341 (0.004)	0.4977 (0.006)
DP-MERF	0.6745 (0.024)	0.4706 (0.019)	0.6849 (0.016)	0.4521 (0.026)
DP-HP	0.6213 (0.072)	0.3478 (0.106)	0.6317 (0.052)	0.2751 (0.124)
Tab-PE	0.6762 (0.017)	0.4049 (0.044)	0.6739 (0.029)	0.4035 (0.063)
Diabetes				
Real	0.6824 (0.001)	0.6536 (0.001)	0.7044 (0.001)	0.6622 (0.001)
BOLT	0.5728 (0.007)	0.6208 (0.006)	0.5808 (0.012)	0.6171 (0.006)
AIM	0.6189 (0.002)	0.6260 (0.007)	0.6235 (0.000)	0.6316 (0.005)
DP-MERF	0.5027 (0.009)	0.6310 (0.000)	0.5000 (0.002)	0.6310 (0.000)
DP-HP	0.4969 (0.017)	0.3758 (0.199)	0.5118 (0.018)	0.2782 (0.164)
Tab-PE	0.5045 (0.019)	0.4866 (0.038)	0.5129 (0.026)	0.4314 (0.172)
MAGIC				
Real	0.8450 (0.000)	0.8461 (0.002)	0.9484 (0.001)	0.9185 (0.003)
BOLT	0.8188 (0.003)	0.8185 (0.005)	0.8169 (0.010)	0.8158 (0.002)
AIM	0.8350 (0.001)	0.8395 (0.005)	0.8735 (0.002)	0.8628 (0.002)
DP-MERF	0.7513 (0.058)	0.8136 (0.029)	0.7401 (0.065)	0.7937 (0.032)
DP-HP	0.7203 (0.057)	0.7720 (0.054)	0.7285 (0.043)	0.7807 (0.047)
Tab-PE	0.7878 (0.006)	0.7839 (0.000)	0.8522 (0.007)	0.8448 (0.009)
Shoppers				
Real	0.8408 (0.003)	0.5489 (0.012)	0.9181 (0.002)	0.6452 (0.008)
BOLT	0.7162 (0.011)	0.3517 (0.009)	0.7707 (0.047)	0.3955 (0.045)
AIM	0.6886 (0.018)	0.3219 (0.033)	0.6861 (0.019)	0.3001 (0.028)
DP-MERF	0.6170 (0.021)	0.2100 (0.049)	0.7386 (0.034)	0.3632 (0.034)
DP-HP	0.6070 (0.058)	0.1515 (0.044)	0.7087 (0.161)	0.2743 (0.208)
Tab-PE	0.6081 (0.049)	0.2467 (0.039)	0.6480 (0.054)	0.2183 (0.059)

Table 7: Dataset-level predictive utility at $\varepsilon = 2$. Entries are arithmetic means (sample standard deviations) across independent releases; observations per cell: Real: 5; BOLT: 5; AIM: 3; DP-MERF: 5; DP-HP: 5; Tab-PE: 5. Real is repeated across privacy-budget tables. LR denotes logistic regression. F1 is positive-class F1; standard deviations are descriptive, not confidence intervals.

Method	LR AUROC	LR F1	LightGBM AUROC	LightGBM F1
Adult				
Real	0.9048 (0.000)	0.6831 (0.001)	0.9265 (0.000)	0.7193 (0.002)
BOLT	0.8726 (0.001)	0.6367 (0.004)	0.8674 (0.002)	0.6225 (0.008)
AIM	0.8880 (0.006)	0.6368 (0.014)	0.8950 (0.004)	0.6694 (0.001)
DP-MERF	0.8158 (0.011)	0.5029 (0.050)	0.7650 (0.076)	0.3659 (0.035)
DP-HP	0.7221 (0.014)	0.4811 (0.016)	0.6714 (0.077)	0.4454 (0.061)
Tab-PE	0.8538 (0.006)	0.5854 (0.008)	0.8717 (0.002)	0.6323 (0.003)
Default				
Real	0.7480 (0.001)	0.5117 (0.004)	0.7744 (0.002)	0.5256 (0.006)
BOLT	0.7305 (0.003)	0.4727 (0.008)	0.7199 (0.004)	0.4722 (0.008)
AIM	0.7369 (0.002)	0.5031 (0.008)	0.7335 (0.002)	0.5029 (0.008)
DP-MERF	0.6708 (0.021)	0.4624 (0.013)	0.6835 (0.021)	0.4449 (0.033)
DP-HP	0.6358 (0.045)	0.4044 (0.043)	0.6421 (0.045)	0.3968 (0.029)
Tab-PE	0.6883 (0.018)	0.4473 (0.017)	0.6884 (0.025)	0.4046 (0.091)
Diabetes				
Real	0.6824 (0.001)	0.6536 (0.001)	0.7044 (0.001)	0.6622 (0.001)
BOLT	0.5840 (0.010)	0.6241 (0.005)	0.5900 (0.008)	0.6147 (0.009)
AIM	0.6097 (0.002)	0.6419 (0.001)	0.6183 (0.005)	0.6412 (0.002)
DP-MERF	0.5023 (0.007)	0.6310 (0.000)	0.4985 (0.015)	0.5104 (0.269)
DP-HP	0.5114 (0.014)	0.6198 (0.014)	0.5188 (0.011)	0.6161 (0.011)
Tab-PE	0.5049 (0.016)	0.4746 (0.048)	0.5231 (0.021)	0.3951 (0.075)
MAGIC				
Real	0.8450 (0.000)	0.8461 (0.002)	0.9484 (0.001)	0.9185 (0.003)
BOLT	0.8211 (0.001)	0.8199 (0.002)	0.8142 (0.005)	0.8143 (0.002)
AIM	0.8397 (0.002)	0.8409 (0.004)	0.8772 (0.000)	0.8683 (0.002)
DP-MERF	0.7554 (0.051)	0.8215 (0.012)	0.7548 (0.046)	0.8022 (0.015)
DP-HP	0.7383 (0.047)	0.7819 (0.040)	0.7541 (0.026)	0.7925 (0.029)
Tab-PE	0.7909 (0.010)	0.7865 (0.006)	0.8673 (0.008)	0.8515 (0.009)
Shoppers				
Real	0.8408 (0.003)	0.5489 (0.012)	0.9181 (0.002)	0.6452 (0.008)
BOLT	0.7210 (0.008)	0.3577 (0.004)	0.7431 (0.026)	0.3650 (0.041)
AIM	0.7465 (0.021)	0.3631 (0.027)	0.7197 (0.011)	0.3518 (0.016)
DP-MERF	0.6427 (0.020)	0.2451 (0.024)	0.7744 (0.025)	0.4078 (0.035)
DP-HP	0.6201 (0.041)	0.2672 (0.047)	0.6598 (0.165)	0.3322 (0.161)
Tab-PE	0.6208 (0.064)	0.2536 (0.069)	0.7005 (0.016)	0.2405 (0.051)

Table 8: Dataset-level predictive utility at $\varepsilon = 8$. Entries are arithmetic means (sample standard deviations) across independent releases; observations per cell: Real: 5; BOLT: 5; AIM: 3; DP-MERF: 5; DP-HP: 5; Tab-PE: 5. Real is repeated across privacy-budget tables. LR denotes logistic regression. F1 is positive-class F1; standard deviations are descriptive, not confidence intervals.

Method	LR AUROC	LR F1	LightGBM AUROC	LightGBM F1
Adult				
Real	0.9048 (0.000)	0.6831 (0.001)	0.9265 (0.000)	0.7193 (0.002)
BOLT	0.8717 (0.002)	0.6370 (0.002)	0.8685 (0.002)	0.6270 (0.004)
AIM	0.9011 (0.002)	0.5890 (0.023)	0.9056 (0.001)	0.6829 (0.002)
DP-MERF	0.8207 (0.016)	0.4916 (0.057)	0.8024 (0.044)	0.3855 (0.059)
DP-HP	0.7345 (0.034)	0.4946 (0.039)	0.7063 (0.041)	0.4734 (0.050)
Tab-PE	0.8774 (0.001)	0.6381 (0.003)	0.8875 (0.002)	0.6530 (0.005)
Default				
Real	0.7480 (0.001)	0.5117 (0.004)	0.7744 (0.002)	0.5256 (0.006)
BOLT	0.7353 (0.001)	0.4741 (0.002)	0.7272 (0.003)	0.4803 (0.002)
AIM	0.7419 (0.002)	0.5050 (0.007)	0.7427 (0.003)	0.4972 (0.003)
DP-MERF	0.6713 (0.020)	0.4552 (0.015)	0.6775 (0.019)	0.4431 (0.025)
DP-HP	0.6535 (0.048)	0.4209 (0.036)	0.6650 (0.020)	0.4005 (0.036)
Tab-PE	0.7162 (0.003)	0.4815 (0.021)	0.7076 (0.009)	0.4315 (0.044)
Diabetes				
Real	0.6824 (0.001)	0.6536 (0.001)	0.7044 (0.001)	0.6622 (0.001)
BOLT	0.5859 (0.003)	0.6290 (0.003)	0.5841 (0.007)	0.6230 (0.006)
AIM	0.6251 (0.024)	0.6369 (0.005)	0.6319 (0.022)	0.6449 (0.002)
DP-MERF	0.5077 (0.011)	0.6310 (0.000)	0.5072 (0.007)	0.6305 (0.001)
DP-HP	0.5295 (0.015)	0.6290 (0.004)	0.5231 (0.008)	0.6283 (0.006)
Tab-PE	0.5185 (0.015)	0.4727 (0.055)	0.5214 (0.021)	0.4247 (0.107)
MAGIC				
Real	0.8450 (0.000)	0.8461 (0.002)	0.9484 (0.001)	0.9185 (0.003)
BOLT	0.8212 (0.001)	0.8213 (0.001)	0.8135 (0.006)	0.8179 (0.002)
AIM	0.8415 (0.000)	0.8421 (0.000)	0.8816 (0.003)	0.8709 (0.005)
DP-MERF	0.7583 (0.053)	0.8198 (0.014)	0.7487 (0.055)	0.8031 (0.025)
DP-HP	0.7346 (0.057)	0.7721 (0.059)	0.7491 (0.044)	0.7820 (0.030)
Tab-PE	0.7926 (0.004)	0.8042 (0.005)	0.8905 (0.005)	0.8691 (0.004)
Shoppers				
Real	0.8408 (0.003)	0.5489 (0.012)	0.9181 (0.002)	0.6452 (0.008)
BOLT	0.7278 (0.008)	0.3597 (0.015)	0.7160 (0.013)	0.3586 (0.017)
AIM	0.7573 (0.008)	0.3904 (0.013)	0.7002 (0.010)	0.3425 (0.011)
DP-MERF	0.7133 (0.019)	0.3221 (0.026)	0.8147 (0.031)	0.4536 (0.062)
DP-HP	0.6722 (0.016)	0.3286 (0.017)	0.7844 (0.072)	0.4207 (0.084)
Tab-PE	0.6574 (0.017)	0.2990 (0.034)	0.7397 (0.031)	0.2582 (0.029)

Table 9: Component-experiment utility at $\varepsilon = 1$. Entries are real-test means $\pm$ sample standard deviations across 5 independent releases per dataset and configuration. Scores lie on $[0, 1]$. All variants use classifier labeling except "512 (latent)", which retains mixture-component labels. The fresh "512 (full)" control shares its private summary and generated covariates with "512 (latent)" and is distinct from the main-grid observations.

Dataset	Conjunctions	Logistic regression		LightGBM	
		AUROC	F1	AUROC	F1
Adult	0	0.7722 ± 0.0060	0.5040 ± 0.0096	0.7498 ± 0.0132	0.4779 ± 0.0216
	128	0.8603 ± 0.0016	0.6279 ± 0.0019	0.8569 ± 0.0031	0.6215 ± 0.0051
	512 (full)	0.8719 ± 0.0025	0.6364 ± 0.0041	0.8644 ± 0.0046	0.6159 ± 0.0114
	2,048	0.8675 ± 0.0005	0.6181 ± 0.0022	0.8652 ± 0.0021	0.6162 ± 0.0027
	512 (latent)	0.8817 ± 0.0041	0.6492 ± 0.0054	0.6012 ± 0.1601	0.2709 ± 0.1871
Default	0	0.6467 ± 0.0286	0.3789 ± 0.0088	0.5252 ± 0.0159	0.3648 ± 0.0075
	128	0.6907 ± 0.0035	0.4385 ± 0.0074	0.6683 ± 0.0156	0.4175 ± 0.0198
	512 (full)	0.7232 ± 0.0038	0.4596 ± 0.0077	0.7064 ± 0.0238	0.4584 ± 0.0163
	2,048	0.7205 ± 0.0057	0.4545 ± 0.0095	0.7144 ± 0.0119	0.4474 ± 0.0152
	512 (latent)	0.7370 ± 0.0030	0.4889 ± 0.0097	0.5401 ± 0.0433	0.3537 ± 0.0252
Diabetes	0	0.5651 ± 0.0026	0.6313 ± 0.0006	0.5661 ± 0.0093	0.6296 ± 0.0009
	128	0.5695 ± 0.0069	0.6300 ± 0.0017	0.5768 ± 0.0095	0.6262 ± 0.0025
	512 (full)	0.5744 ± 0.0065	0.6197 ± 0.0051	0.5801 ± 0.0063	0.6147 ± 0.0080
	2,048	0.5918 ± 0.0082	0.5731 ± 0.0096	0.5968 ± 0.0064	0.5799 ± 0.0074
	512 (latent)	0.5976 ± 0.0044	0.5497 ± 0.0149	0.6343 ± 0.0280	0.4690 ± 0.1531
MAGIC	0	0.8220 ± 0.0014	0.8289 ± 0.0026	0.8212 ± 0.0023	0.8239 ± 0.0043
	128	0.8218 ± 0.0009	0.8217 ± 0.0034	0.8207 ± 0.0043	0.8146 ± 0.0014
	512 (full)	0.8194 ± 0.0022	0.8193 ± 0.0031	0.8173 ± 0.0084	0.8166 ± 0.0014
	2,048	0.8199 ± 0.0020	0.8248 ± 0.0032	0.8102 ± 0.0072	0.8238 ± 0.0038
	512 (latent)	0.8259 ± 0.0026	0.8338 ± 0.0079	0.8124 ± 0.0171	0.8303 ± 0.0074
Shoppers	0	0.7031 ± 0.0396	0.3028 ± 0.0223	0.6244 ± 0.1137	0.2587 ± 0.0206
	128	0.6609 ± 0.0196	0.2845 ± 0.0211	0.6585 ± 0.0311	0.2816 ± 0.0331
	512 (full)	0.7135 ± 0.0139	0.3426 ± 0.0190	0.7543 ± 0.0464	0.3929 ± 0.0474
	2,048	0.7221 ± 0.0149	0.3190 ± 0.0235	0.7202 ± 0.0142	0.2837 ± 0.0405
	512 (latent)	0.6928 ± 0.0048	0.3113 ± 0.0210	0.7260 ± 0.1606	0.4793 ± 0.2062

D.2 Conjunction count and paired labeling effects

The full component control is the paired classifier-label experiment. It is separate from the main-grid observations. Equal-dataset effects, in LR AUROC/LR F1/LightGBM AUROC/LightGBM F1 order, are: 512 versus 0: $+3.866/+4.633/+8.717/+6.870$ percentage points; 512 versus 128: $+1.985/+1.502/+2.826/+2.741$ percentage points; 2,048 versus 512: $+0.388/-1.763/-0.313/-2.950$ percentage points. Changing conjunction count changes both the feature workload and its calibrated measurement noise.

Table 10: Effect of classifier labeling: classifier labels minus retained mixture-component labels, in percentage points. Each dataset entry is the mean ± sample standard deviation of 5 paired score differences. Positive values favor classifier labeling. Pairs share the private summary, generated covariates, and evaluation seed; label-stratified evaluation partitions can differ. The final row averages dataset means equally and has no omnibus standard deviation.

Dataset	Logistic regression		LightGBM	
	AUROC	F1	AUROC	F1
Adult	-0.97 ± 0.31	-1.28 ± 0.52	$+26.32 \pm 15.99$	$+34.50 \pm 17.71$
Default	-1.37 ± 0.30	-2.93 ± 1.55	$+16.62 \pm 1.99$	$+10.48 \pm 2.04$
Diabetes	-2.32 ± 1.02	$+7.00 \pm 1.94$	-5.42 ± 2.78	$+14.57 \pm 15.04$
MAGIC	-0.65 ± 0.27	-1.46 ± 1.09	$+0.50 \pm 0.89$	-1.37 ± 0.83
Shoppers	$+2.07 \pm 1.69$	$+3.13 \pm 3.29$	$+2.83 \pm 12.79$	-8.64 ± 17.78
Equal-dataset mean	-0.65	$+0.89$	$+8.17$	$+9.91$

Table 11: Absolute complete-table synthesis time in seconds. Releases are averaged within each dataset–privacy cell, then datasets and privacy levels receive equal weights. Blue and lilac mark the smallest and second-smallest entries using unrounded values. Rows are ordered by the final column. Times reflect the recorded implementations and hardware (Appendix C.6).

Method	$\varepsilon = 0.5$	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 8$	All ε
BOLT	**2.88**	**2.87**	**2.83**	**2.83**	**2.85**
DP-MERF	**37.81**	**37.84**	**37.28**	**34.56**	**36.87**
DP-HP	718.44	723.90	723.53	713.61	719.87
Tab-PE	905.89	899.75	938.13	919.52	915.82
AIM	1,329.17	3,534.00	31,368.43	38,116.74	18,587.09

D.3 Absolute synthesis time

Absolute times and headline runtime ratios use the same arithmetic means. BOLT averages 2.852 seconds across equally weighted dataset–privacy cells. Comparator/BOLT ratios are DP-MERF 12.9×; DP-HP 252.4×; Tab-PE 321.1×; AIM 6,516.7×. These overall ratios need not describe each privacy budget separately.