Leakage-Aware LLM Augmentation for Attrition Prediction: A Decision-Centric Evaluation

Liao Weiquan[1], Jiayangmei Xu[2] & Ekaterina A. Panova[3,4]

[1] Shenzhen MSU-BIT University, China
[2] Beihang University, China
[3] School of Management, Shenzhen MSU-BIT University, China
[4] School of Public Administration, Lomonosov Moscow State University, Moscow, Russian Federation

Correspondence: Ekaterina A. Panova, School of Public Administration, Lomonosov Moscow State University, Moscow, Russian Federation.

Received: August 15, 2026 Accepted: September 19, 2026 Published: September 23, 2026
doi:10.65343/aiis.v2i2.141 URL: https://doi.org/10.65343/aiis.v2i2.141

Abstract

Employee attrition is a high-cost, asymmetric decision problem plagued by data imbalance. To address this, we propose a leakage-aware, dual-network augmentation framework that integrates the semantic reasoning of Large Language Models (LLMs) with the distributional rigorousness of GAN discriminators. Specifically, we employ an autoregressive transformer (GReaT) as a generative "actor" to synthesize diverse minority samples, coupled with a post-hoc "critic" derived from a conditional GAN to strictly filter low-fidelity outliers. This actor-critic inspired loop ensures that synthetic records broaden coverage without introducing noise. We evaluate this pipeline using a rigorous protocol where all generation and filtering occur strictly within training folds to prevent leakage. Experiments on HR datasets show that moderate, critic-guided oversampling yields significant recall gains (e.g., raising recall from 45% to 53%) and improved F_1 scores compared to baselines, while maintaining calibration and discrimination (stable ROC-AUC). Furthermore, the approach preserves interpretability (SHAP) and fairness (low TPR gaps), offering HR practitioners a robust, decision-centric tool to identify at-risk employees without sacrificing transparency.

Keywords: employee attrition, HR analytics, large language models, actor-critic framework, synthetic data augmentation, leakage-aware evaluation, class imbalance

1. Introduction

Employee attrition creates a decision problem with asymmetric costs. HR leaders must allocate limited retention budgets under uncertainty and trade off missed preventable departures against avoidable interventions. Credible industry evidence estimates replacement costs at about $0.5–2.0 \times$ *annual salary* per role, totaling roughly **USD 1 trillion** per year in the U.S. (Harter, & Adkins, 2019). We therefore frame attrition prediction as *decision support*. A useful model prioritizes the minority class (*Attrition=Yes*) at tolerable false-alarm rates, produces interpretable outputs that guide action, and supports equitable decisions across subgroups.

A practical barrier is that widely used HR datasets are small and highly imbalanced. A common shortcut—"balance then split" (e.g., oversampling before cross-validation or a hold-out split)—leaks target information across folds and biases estimates, especially for small tabular data (Demircioğlu, 2024). Reproducibility checklists in leading venues also call for explicit disclosure of data handling and leakage controls (NeurIPS, 2025). These points motivate *leakage-aware* evaluation for any augmentation that seeks to improve minority recall and the precision–recall trade-off.

Recent progress in *LLM-based augmentation* offers a principled augmentation route. GReaT encodes rows as text, fine-tunes an autoregressive transformer, and supports *arbitrary conditioning* to sample realistic rows that capture cross-feature dependencies (Borisov, Seßler, Leemann, Pawelczyk, & Kasneci, 2023). We apply conditional sampling strictly *within* training folds to synthesize minority instances without breaking out-of-fold integrity, aiming to improve decision-relevant metrics.

2. Related Work

We review methods for class imbalance, LLM-based augmentation, and standards for interpretability, fairness, and rigorous evaluation. Classical remedies such as SMOTE/ADASYN and tabular GANs appear frequently in HR analytics (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Salimans, Goodfellow, Zaremba, Cheung, Radford, & Chen, 2016; Turner, Hung, Frank, Saatchi, & Yosinski, 2019); however, unless resampling remains strictly in-

fold, these methods induce optimistic bias via leakage (Demircioğlu, 2024). In contrast, LLM-based augmentation (e.g., GReaT) model joint dependencies over textual encodings and enable *arbitrary conditioning* without bespoke retraining, making them suitable for principled augmentation when applied in-fold (Borisov, Seßler, Leemann, Pawelczyk, & Kasneci, 2023).

For transparency, SHAP is a common choice to attribute predictions to features and to translate them into actionable managerial levers (Lundberg, & Lee, 2017). To assess equity, we follow the *equal opportunity* criterion and report true positive rate (TPR) gaps across demographic subgroups (Hardt, Price, & Srebro, 2016). For rigorous comparison, we pair ROC-AUC/PR-AUC with statistical tests for AUC differences (e.g., DeLong) when appropriate (Molodianovitch, Faraggi, & Reiser, 2006).

3. Research Gap

From a decision perspective, the literature lacks:

1. Leakage-free, decision-relevant evaluations of LLM-based augmentation on highly imbalanced HR data, with metrics and summaries aligned to asymmetric costs and budget constraints;
2. Head-to-head evidence on how *classical ML* versus a compact *transformer–MLP* exploit synthetic minority samples under identical, leakage-controlled protocols;
3. Demonstrations that numeric gains translate into interpretable, actionable, and fair decisions (higher minority recall at tolerable false-alarm rates, SHAP levers mapping to HR actions, and bounded subgroup TPR gaps);
4. Systematic analysis of *synthetic scale effects* (progressive augmentation) to guide budgeted deployment.

To overcome the limitations of single-model augmentation, we propose a dual-network generative framework that integrates the semantic reasoning of Large Language Models (LLMs) with the distributional strictness of GAN discriminators. In this Actor-Critic inspired architecture, the LLM serves as the 'Actor,' utilizing its pre-trained knowledge to explore the diverse semantic space of employee profiles and generate candidate samples with rich feature dependencies. Conversely, the pre-trained GAN discriminator acts as the 'Critic,' providing a rigorous feedback signal based on the learned manifold of the real data. This hybrid approach essentially establishes a cybernetic quality control loop: the Actor ensures diversity and logic, while the Critic imposes distributional fidelity, filtering out 'hallucinated' outliers that statistically deviate from the true data distribution. This synergy allows us to synthesize high-utility minority samples that are both realistic and strictly aligned with the underlying data structure.

4. Our Contributions

- **Decision-centric LLM augmentation.** We adapt GReaT-style conditional generation (Borisov, Seßler, Leemann, Pawelczyk, & Kasneci, 2023) to synthesize minority (*Attrition=Yes*) records strictly *within* training folds, select models on validation folds, and leave test folds untouched, with the goal of reducing costly false negatives.
- **Leakage-controlled evaluation.** We use a repeated, stratified hold-out/cross-validation design and keep all fine-tuning and synthetic generation in-fold, addressing the oversampling-before-split bias (Demircioğlu, 2024) and aligning with checklist expectations on data handling and reproducibility (NeurIPS, 2025).
- **Dual benchmark suite and statistics.** We compare logistic and boosted-tree baselines to a compact transformer–MLP; we report ROC-AUC, PR-AUC, balanced accuracy, and decision-facing summaries (alerts per true save), and apply DeLong tests for AUC differences across seeds.
- **Interpretability and fairness.** We provide SHAP explanations as retention levers (Lundberg, & Lee, 2017) and monitor equal-opportunity gaps (TPR parity) across demographic subgroups (Hardt, Price, & Srebro, 2016).
- **Progressive augmentation guidance.** We trace performance and intervention load across synthetic scales (e.g., $0.25\times$ to $4\times$ the minority count) and identify robust operating regions for practice.
- **Open artifacts.** We release code, prompts, and configs to support reproduction consistent with community checklists (NeurIPS, 2025).

5. Materials and Methods

5.1 Dataset and Preprocessing

We primarily use the IBM HR Attrition dataset (about 1.5k employees) containing demographic, performance, and workplace features along with a binary attrition label ("Yes"/"No"). The minority ("Yes") rate is approximately 15–17%, which makes both learning and evaluation non-trivial (IBM Watson Analytics, 2015).

To further validate the robustness and generalizability of our approach, we additionally conduct experiments on a larger HR dataset with around 15k employee records, constructed with similar feature structures and label distributions. This supplementary dataset allows us to examine whether the observed performance trends hold

under larger-scale conditions.

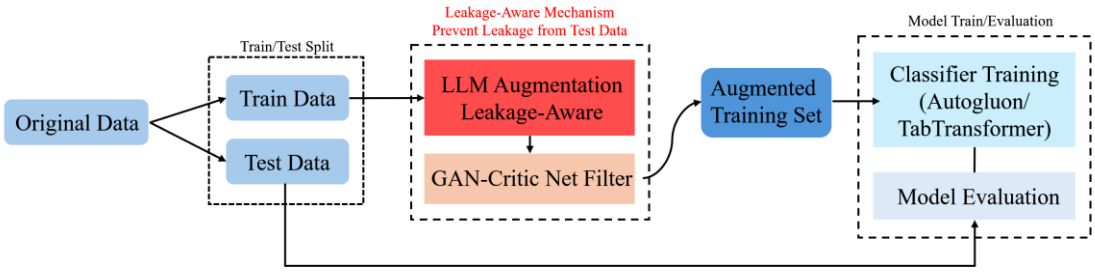


Figure 1. Leakage-Aware LLM Augmentation Framework

As Figure 1 shows, we avoid leakage by fitting all preprocessing on the training split only and then applying it to validation/test. We standardize numerical features and one-hot encode categorical features using a standard transformer stack (Fabian, 2011). We adopt a stratified 70/30 train–test split with a fixed random seed to preserve class ratios across runs (Fabian, 2011).

To illustrate class overlap, we project the training set into two dimensions using PCA and t-SNE. PCA gives a linear projection that retains maximum variance; t-SNE provides a non-linear, neighborhood-preserving view for exploration (Fabian, 2011; Jolliffe, & Cadima, 2016; van der Maaten, & Hinton, 2008, November). As shown in Figure 2, attriting and non-attriting employees intermingle instead of forming a distinct cluster. Single drivers (e.g., OverTime, JobSatisfaction) cannot fully separate the classes, which suggests that higher-capacity models and augmentation may help.

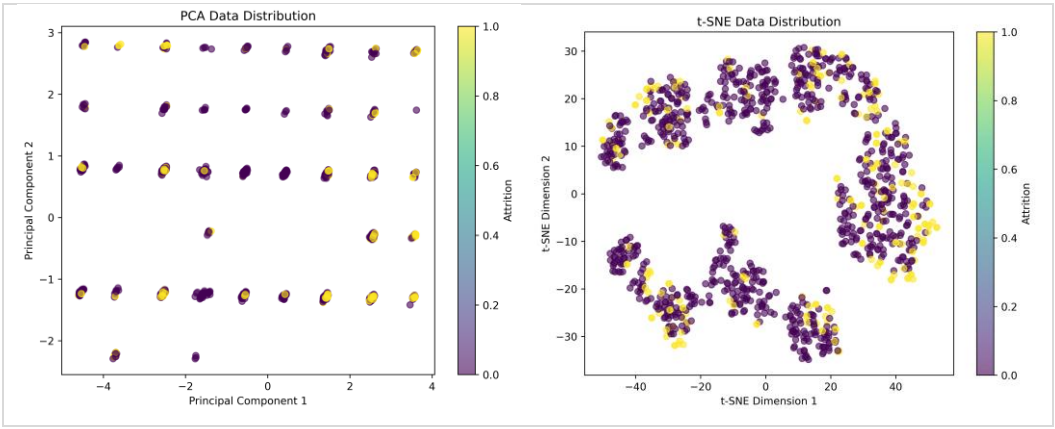


Figure 2. Visualization of the training data in (left) PCA and (right) t-SNE space. Purple points denote non-attrition and yellow points denote attrition. The considerable overlap highlights the challenge

5.2 Synthetic Data Generation

To address class imbalance, we synthesize minority (`Attrition=Yes`) samples via **LLM-based augmentation**. We adopt **GReaT** (Generation of Realistic Tabular data) (Borisov, Seßler, Leemann, Pawelczyk, & Kasneci, 2023), which fine-tunes an autoregressive LLM on text-encoded rows and supports *fully-conditional* sampling. All fitting and sampling occur strictly *within* training folds to keep evaluation leakage-free.

5.2.1 LLM Setup

We fine-tune a distilled transformer (~1.5B parameters) on the training split's positive cases using LoRA adapters (low-rank updates with frozen base weights) to cut trainable parameters and memory (Hu, *et al.*, 2022). After up to 100 epochs, we sample additional "Yes" instances. For diversity and control, we enable *guided sampling* (feature-by-feature generation) with a random feature order and temperature 0.7, which GReaT recommends for wide, mixed-type tables.

5.2.2 Scale of Augmentation and Validity Checks

We add $N = 50$ and $N = 100$ synthetic "Yes" records (about 25% and 50% oversampling of the minority

class) to the original training data. We apply schema/range checks (e.g., nonnegative age and working years), de-duplicate near-duplicates, and drop any constraint violations before training.

5.2.3 CTGAN Baseline

For comparison, we train a CTGAN baseline on the same in-fold training data using *ydata-synthetic* (Barr, Rozman, & Guo, 2025; Ding, Wang, Wang, & Welch, 2023; YData, 2025). We set epochs= 500, latent/embedding= 128, generator dims= $(256,256)$, critic dims= $(256,256)$, PAC= 10, batch= 64, and Adam with learning rate 2×10^{-4} and $(\beta_1, \beta_2) = (0.5, 0.9)$. CTGAN uses conditional vectors and training-by-sampling for imbalanced discrete columns and mode-specific normalization for non-Gaussian continuous features (Xu, Skoularidou, Cuesta-Infante, & Veeramachaneni, 2019). We monitor generator/critic losses, checkpoint the best model, and use conservative early stopping. Relative to CTGAN, the GReaT pipeline is operationally simpler (no adversarial dynamics or heavy retuning) while retaining fully-conditional sampling.

5.3 Predictive Models and Training

We evaluate five classifiers:

1. **Logistic Regression (LR).** Linear baseline with L_2 regularization and class weighting (Fabian, 2011).

2. **Random Forest (RF).** An ensemble of trees ($n_{\text{trees}} = 100$) with balanced in-bag subsampling to address imbalance (balanced random-forest style) (Fabian, 2011; Chen, Liaw, Breiman, *et al.*, 2004).

3. **XGBoost.** Gradient-boosted trees with depth and learning rate tuned via cross-validation; we set *scale_pos_weight* for imbalance (Chen, & Guestrin, 2016; XGBoost Developers, 2025).

4. **TabTransformer.** A transformer-based model that learns contextual embeddings for categorical features; we follow Huang *et al.* with three self-attention layers and train for 40 epochs using binary cross-entropy (Huang, Khetan, Cvitkovic, & Karnin, 2020).

5. **AutoGluon-Tabular.** An AutoML stack/ensemble (trees and neural nets) trained in multiple layers; we use the *best_quality* or *extreme_quality* preset on GPU (Erickson, *et al.*, 2020; Tang, *et al.*, 2024).

Dataset $\mathcal{D} = \{(X, y)\}$, Fold count K, Classifier $\mathcal{C}$, Augmentation $\mathcal{A}$ Performance Metrics vector $\mathcal{M}_{final}$ (e.g., AUC, Recall)

Initialize: $\mathcal{L} \leftarrow []$ **Partition:** Split $\mathcal{D}$ into K stratified folds $\{F_1, F_2, \ldots, F_K\}$

$\mathcal{D}_{test} \leftarrow F_k$ $\mathcal{D}_{train} \leftarrow \mathcal{D} \backslash F_k$

Fit transformations Φ (scaler/encoder) on $\mathcal{D}_{train}$ Apply Φ: $\mathcal{D}_{train} \leftarrow \Phi(\mathcal{D}_{train})$, $\mathcal{D}_{test} \leftarrow \Phi(\mathcal{D}_{test})$

Identify minority subset $\mathcal{D}_{min} \subset \mathcal{D}_{train}$ $\mathcal{S} \leftarrow \mathcal{A}(\mathcal{D}_{min}, \Phi)$ $\mathcal{D}_{augmented} \leftarrow \mathcal{D}_{train} \cup \mathcal{S}$

Train classifier $\mathcal{C}$ on $\mathcal{D}_{augmented}$ Evaluate $\hat{y} \leftarrow \mathcal{C}(\mathcal{D}_{test})$ Record metrics to $\mathcal{L}$

Mean and Standard Deviation of $\mathcal{L}$

All models use the same training split. Baselines train on original data; augmentation runs refit on the augmented set (original + synthetic). To ensure the rigorousness of our evaluation and strictly prevent data leakage, we formalized the training procedure in Algorithm [algo: leakage_aware]. This protocol guarantees that all learnable components—including feature scalers, encoders, and synthetic data generators—are fitted solely on the training folds, preserving the integrity of the hold-out test sets. For LR we set *class_weight='balanced'*; for XGBoost we tune *scale_pos_weight*; for RF we downsample the majority class within each bootstrap draw. We select hyperparameters (e.g., LR regularization, XGBoost depth/learning rate) by grid search with 5-fold *stratified* cross-validation optimizing ROC-AUC; AutoGluon performs its own internal tuning. No model accesses any information about the "Attrition" label beyond its training data (Fabian, 2011).

5.4 Evaluation Metrics

We evaluate discrimination, fairness, calibration, interpretability, and augmentation scalability.

- **Discrimination.** We report ROC-AUC and Average Precision (AP; area under the Precision–Recall curve (Flach, & Kull, 2015)) on the test set and plot ROC and PR curves (Figure 3(b)). PR curves emphasize minority-class performance when positives are rare (Davis, & Goadrich, 2006; Saito, & Rehmsmeier, 2015). We also show a confusion matrix for the best model (Figure 4(a)) and derive precision and recall for the attrition class. In PR space, the no-skill baseline is a horizontal line at the positive prevalence (here $\approx 15\%$), and AP measures improvement over this baseline.

- **Fairness.** Following *equal opportunity*, we compute the true positive rate (TPR) for females and males and report the *TPR gap* $\Delta\text{TPR} = |\text{TPR}_{\text{male}} - \text{TPR}_{\text{female}}|$ on the test set (Hardt, Price, & Srebro, 2016). We do not impose fairness constraints; this diagnostic surfaces potential bias. Because base attrition rates differ slightly by gender, we report groupwise TPRs alongside the gap.

- **Calibration.** We assess probability calibration with reliability diagrams (Figure 4 (b)) and report the Brier score (mean squared error between predicted probabilities and outcomes; lower is better) (Glenn, et al., 1950; Silva Filho, Song, Perello-Nieto, Santos-Rodriguez, Kull, & Flach, 2023).
- **Interpretability.** We use SHAP to attribute predictions and summarize global importance for the top model (AutoGluon ensemble) (Lundberg, & Lee, 2017). A SHAP summary plot (Figure 6) ranks features and visualizes value distributions, helping verify alignment with HR factors (e.g., *OverTime*, *Age*, *MonthlyIncome*, *JobSatisfaction*).
- **Scalability and augmentation impact.** Starting from the baseline, we evaluate after adding $+50$ and $+100$ synthetic "Yes" examples (training size increases of roughly 17% and 34%). We track ROC-AUC, AP, fairness (TPR gap), and calibration (Brier), and visualize PR curves for AutoGluon and TabTransformer (Figure (5b)). We also overlay real vs. synthetic projections (PCA/t-SNE) to monitor drift. In our runs, GReaT samples followed the real-data structure and preserved salient correlations, while CTGAN required careful tuning to avoid artifacts (e.g., duplicates or out-of-range values), consistent with prior reports (Xu, Skoularidou, Cuesta-Infante, & Veeramachaneni, 2019).

We compared ROC–AUC across $K = 6$ models, and six AutoGluon variants trained with different augmentation settings (LLM+50/100/200 and GAN+50/100/200), all evaluated on the same test set using DeLong's variance–covariance estimator. First, we ran an omnibus Wald χ^2 test with $df = K - 1$ to assess whether all AUCs are equal. If significant, we conducted pairwise DeLong tests on AUC differences, reporting $\Delta AUC, z$,and two-sided p.To control multiplicity, we applied Holm (FWER) and Benjamini–Hochberg (BH-FDR) corrections to the pairwise p-values. DeLong 95% CIs for individual AUCs were computed via the normal approximation.

5.5 Critic-Based Filtering for Quality Enhancement

To further improve the quality of LLM-based synthetic data, we introduce a post-hoc *critic-based* filter inspired by the idea of *discriminator rejection sampling* [31]. Specifically, after training a CTGAN on the *training* split, we freeze its discriminator $D_{\text{CTGAN}}(x)$ and use it to score each LLM–GReaT sample x via the logit:

$$s(x) = \log \frac{D_{\text{CTGAN}}(x)}{1 - D_{\text{CTGAN}}(x)},$$

which is proportional to $p_{\text{data}}(x)/p_g(x)$ under the optimal discriminator. (Note 1) We then **rank** all LLM-generated records by $s(x)$ and retain the top-K (or top-$q\%$), discarding low-scoring outliers. This filtering procedure effectively improves data fidelity while maintaining coverage, analogous to the DRS approach for image generation (Azadi, Olsson, Darrell, Goodfellow, & Odena, 2018).

The complete pipeline, integrating the GReaT-based generation with this post-hoc critic filtering, is detailed in Algorithm [algo: critic_gen]. As shown, the process acts as a quality gate: we rank all LLM-generated records by $s(x)$ and retain the top-K samples. By explicitly filtering out low-scoring instances—which typically exhibit significant distributional shifts from the real data—we ensure that the augmentation broadens the minority class coverage without introducing harmful noise or artifacts.

Furthermore, since filtering inherently alters the data distribution, it is also relevant to fairness concerns discussed in works such as (Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 2015), where selection mechanisms can unintentionally induce disparate impact across subgroups.

Minority Data $\mathcal{D}_{pos}$, Augmentation Target N, Filter Top-K High-Fidelity Synthetic Set $\mathcal{S}_{final}$

Fine-tune Transformer $\mathcal{M}_{LLM}$ on serialized $\mathcal{D}_{pos}$ with LoRA Sample candidate set $\mathcal{S}_{raw}$ $(|\mathcal{S}_{raw}| > N)$ via conditional generation Apply schematic constraints (e.g., non-negative checks) to $\mathcal{S}_{raw}$

Train Discriminator D_{critic} on $\mathcal{D}_{train}$ to distinguish Real vs. Fake **Freeze** D_{critic} weights Initialize scores

$$V \leftarrow []$$

Compute quality score $s(x) \leftarrow \log(D_{critic}(x)) - \log(1 - D_{critic}(x))$ Append $(x, s(x))$ to V

Sort V by score descending $\mathcal{S}_{final} \leftarrow$ Select top-K instances from V

$$\mathcal{S}_{final}$$

5.5.1 Design and Safeguards

(i) *Leakage control:* we train the critic strictly in-fold and only score LLM samples from the same fold; (ii) *Calibration (optional):* we can apply temperature scaling on a small validation subset before ranking; (iii) *Budgeted selection:* we set K or q on validation to balance fidelity and diversity.

5.5.2 Relation to Prior Art

This filter follows the idea of using a GAN discriminator for rejection/subsampling (e.g., Discriminator Rejection

Sampling and MH-GAN), adapted here to tabular LLM outputs as a *post-hoc* quality gate. It adds no extra training beyond the CTGAN baseline and is lightweight for practice. We leave a detailed ablation to future work.

6. Results

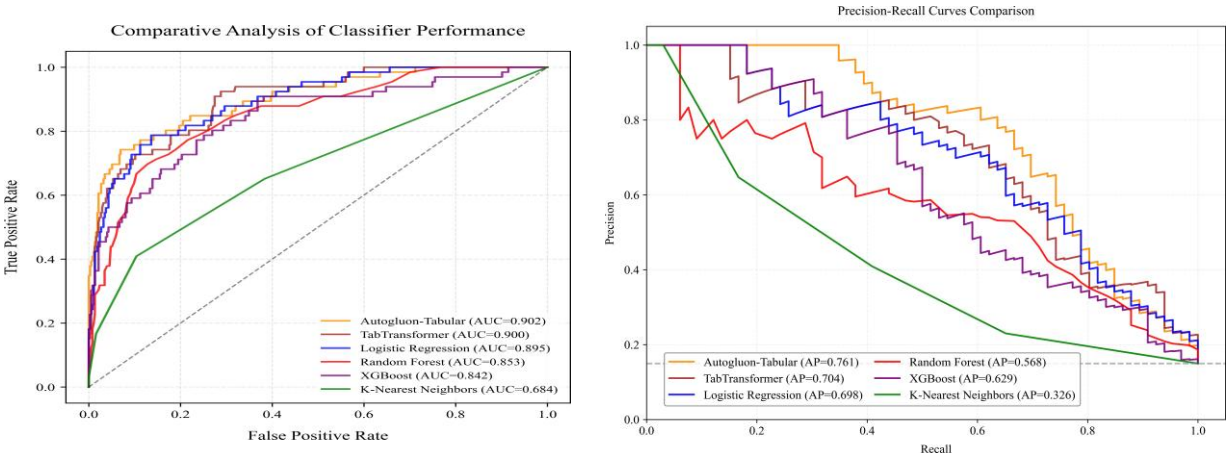
Baseline ROC and Precision–Recall on the test set (no augmentation). AutoGluon-Tabular attains the best discrimination and AP (AUC ≈ 0.88, *AP* $= 0.761$*), followed by TabTransformer (AUC* ≈ 0.85, *AP* $= 0.704$*) and Logistic Regression (AUC* ≈ 0.83, *AP* $= 0.698$*). The dashed PR line is the no-skill precision at the attrition rate (*≈ 0.15*). The omnibus DeLong/Wald test rejects equality of AUCs (*$\chi^2(5) = 45.469,\ p = 1.165 \times 10^{-8}$*). Holm-adjusted post-hoc (FWER) indicates* ***AutoGluon*** $>$ *XGBoost, Random Forest, KNN and TabTransformer, Logistic Regression, XGBoost, Random Forest* $>$ *KNN; all other pairwise differences are n.s. (*$p_{Holm} \geq 0.05$*).*

6.1 Overall Model Performance

Table 1. Overall performance (mean over 20 repeated stratified splits). AutoGluon-Tabular ranks first on ROC-AUC and AP, followed by TabTransformer; tree-based and linear baselines are lower.

Model	**Performance Evaluation**				
2-6	ROC AUC	Precision	Recall	F_1	Avg Precision
AutoGluon	**0.902**	**0.900**	**0.409**	**0.562**	**0.761**
TabTransformer	0.899	0.827	0.363	0.505	0.704
Logistic Regression	0.894	0.833	0.303	0.444	0.698
Random Forest	0.852	0.800	0.181	0.296	0.568
XGBoost	0.841	0.800	0.363	0.500	0.629

Table 1 summarizes test metrics, and Figure 3 shows ROC/PR curves. **AutoGluon-Tabular** performs best (ROC-AUC $= 0.902$, AP $= 0.761$). **TabTransformer** is a close second (ROC-AUC $= 0.899$, AP $= 0.704$). **Logistic Regression** (ROC-AUC $= 0.894$, AP $= 0.698$) and **XGBoost** (ROC-AUC $= 0.841$, AP $= 0.629$) follow. **Random Forest** has a similar ROC-AUC (0.852) but a lower AP (0.568), indicating faster precision decay as recall increases. In short, ensemble/transformer models outperform linear or tree baselines on this task.



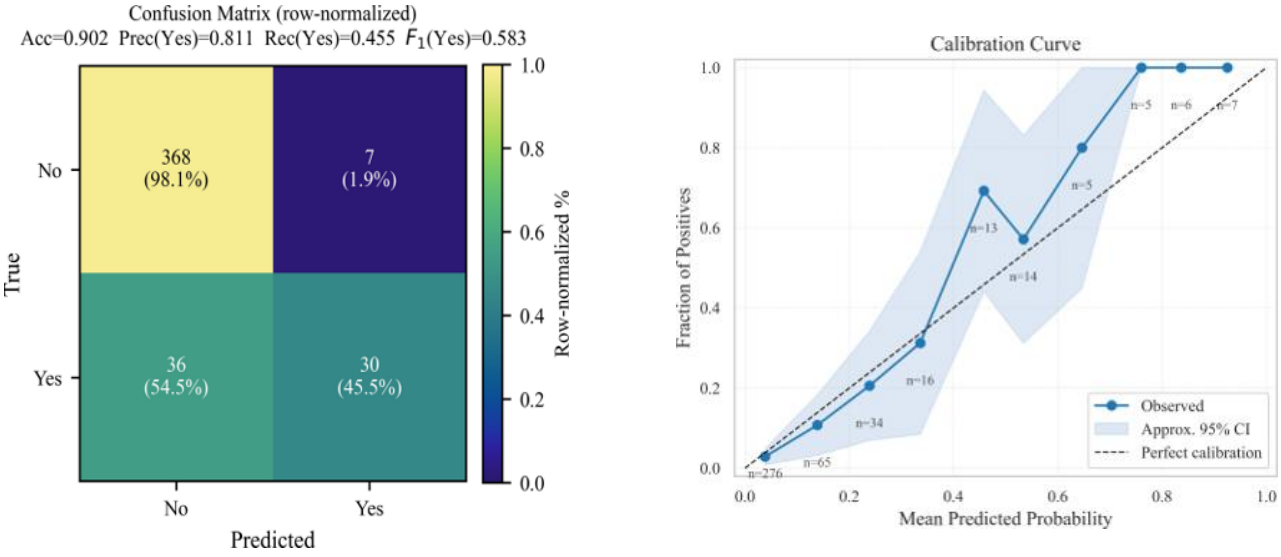
(a) ROC (no augmentation). AutoGluon and TabTransformer lead in the high-recall region

(b) Precision–Recall (no augmentation). AutoGluon sustains higher precision across recall; RF and XGBoost drop faster

Figure 3. Baseline ROC and Precision–Recall on the test set (no augmentation). AutoGluon-Tabular attains the best discrimination and AP (AUC ≈ *0.88, AP = 0.761), followed by TabTransformer (AUC ≈ 0.85, AP = 0.704) and Logistic Regression (AUC ≈ 0.83, AP = 0.698). The dashed PR line is the no-skill precision at the attrition rate (≈ 0.15). The omnibus DeLong/Wald test rejects equality of AUCs (χ2(5) = 45.469, p = 1.165 × 10−8). Holm-adjusted post-hoc (FWER) indicates AutoGluon > {XGBoost, Random Forest, KNN} and*

{TabTransformer, Logistic Regression, XGBoost, Random Forest} > KNN; all other pairwise differences are n.s. (pHolm ≥ 0.05).

To ground the effect, Figure 4(a)) reports the AutoGluon-Tabular confusion matrix. Of 375 non-attrition employees, 368 are true negatives and 7 false positives; among 66 attrition cases, 30 are true positives and 36 false negatives. Thus **accuracy** $= 90.2\%$, **precision** (attrition) $= 81.\%$, and **recall** (attrition) $= 45.5\%$. High precision means few false alarms; moderate recall suggests threshold tuning can trade precision for more coverage of at-risk employees. In a retention context, this means our model is reliable when it does predict attrition but misses over half of eventual leavers at default threshold – a trade-off that HR managers could adjust depending on intervention resources and risk tolerance.



(a) Confusion matrix for AutoGluon-Tabular

(b) Calibration curve (reliability diagram) on the test set. Closer to the diagonal is better; error bars are binomial.

Figure 4. Confusion matrix and calibration of AutoGluon-Tabular, side by side for compact comparison

6.2 Fairness Analysis

We assess gender fairness under *equal opportunity* (groupwise TPR on the attrition class). For AutoGluon-Tabular, **female TPR** is 42.1% and **male TPR** is 50.0%, so the TPR gap is about **8%**. This indicates slightly higher sensitivity for males. Because base rates differ slightly by gender, we track both groupwise TPRs and the gap; we do not impose fairness constraints in training. False positive rates are low (about 2–3%) and similar across genders. From a HR standpoint, this modest gap suggests monitoring is needed to ensure attrition interventions reach all groups fairly, but the bias is not large in absolute terms.

6.3 Reliability and Calibration

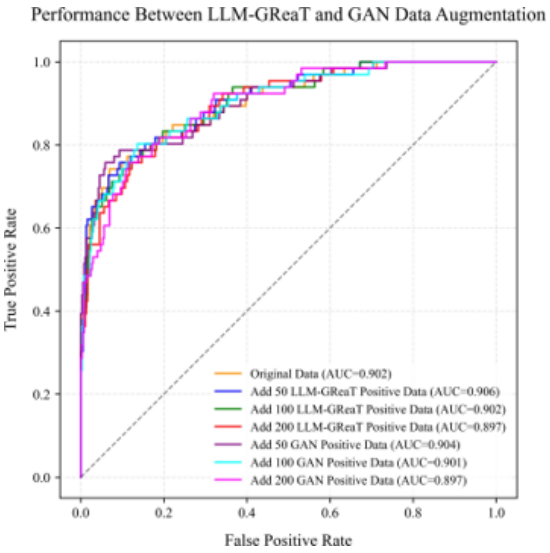
Figure 6 shows a reliability diagram for AutoGluon-Tabular. Predictions are well calibrated at low probabilities, with mild underestimation in the mid-range. The **Brier score** is 0.103, better than logistic regression (0.129). A constant forecaster at the base rate (15%) would achieve $p(1-p)! \approx !0.1275$. We did not apply post-hoc calibration (e.g., Platt scaling or isotonic regression) to keep the evaluation on raw outputs.

6.4 GReaT vs. GAN: Practical Comparison

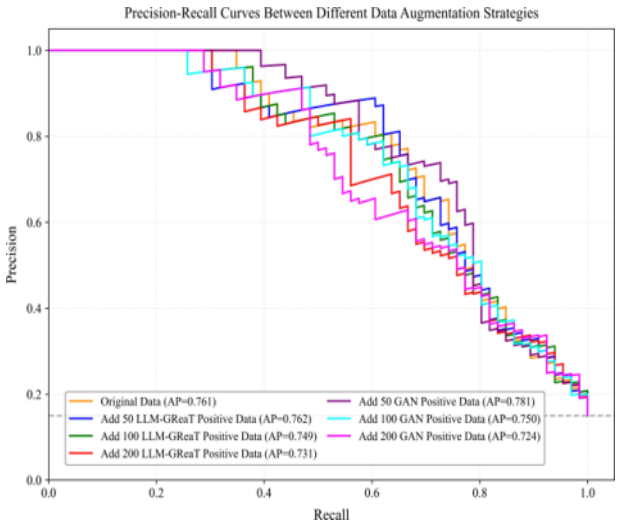
ROC and PR with synthetic augmentation (LLM-based augmentation via GReaT vs. CTGAN). ROC varies little with scale, while PR reveals trade-offs: for AutoGluon, AP decreases slightly with $+50/+100$ *(*$0.761 \rightarrow 0.748 \rightarrow 0.736$*); TabTransformer drops more (*$0.704 \rightarrow 0.605/0.607$*). CTGAN tends to raise precision at small scales at a cost to recall. Over-augmentation degrades PR, highlighting the need for quality control.*

We next compare LLM-based augmentation (GReaT) to CTGAN. Figure 5 contrasts three training sets per model (none, $+50$, $+100$ synthetic "Yes"). For AutoGluon-Tabular, AP declines slightly as synthetic size grows;

TabTransformer drops more. The likely cause is small distributional shifts from synthetic samples. CTGAN shows higher precision at small scales but lower recall. Overall, ROC is stable across scales, whereas PR exposes method-specific trade-offs.



(a) ROC under augmentation. ROC is comparatively stable across LLM-based augmentation (GReaT) and CTGAN



(b) Precision–Recall under augmentation. LLM-based augmentation tends to favor recall/F1 at moderate scale; CTGAN favors precision.

Figure 5. ROC and PR with synthetic augmentation (LLM-based augmentation via GReaT vs. CTGAN). ROC varies little with scale, while PR reveals trade-offs: for AutoGluon, AP decreases slightly with +50/ + 100 (0.761 → 0.748 → 0.736); TabTransformer drops more (0.704 → 0.605/0.607). CTGAN tends to raise precision at small scales at a cost to recall. Over-augmentation degrades PR, highlighting the need for quality control.

6.4. Interpretability of Model Predictions

To ensure transparency, we examine SHAP for AutoGluon-Tabular (Figure 6). OverTime, Busi- nessTravel, NumCompaniesWorked, JobRole, and DistanceFromHome have the largest effects. Signs align with HR intuition: more overtime, frequent travel, many past employers, certain roles, and long commutes raise risk; satisfaction and tenure dynamics moderate it. Sensitive attributes (e.g., gender) are not top drivers, which lowers fairness concerns and yields actionable levers.

6.5. GReaT vs. GAN: Practical Comparison

GReaT fine-tuning is straightforward and stable, yielding diverse, plausible samples. CTGAN is powerful but requires careful tuning and can produce near-duplicates or constraint violations that need filtering. Multiple LLM sampling runs are more consistent than GAN runs in practice. For reuse, LLM-based augmentation adapts with light prompting or brief fine-tuning; GANs typically need full retraining.

Table 2 shows two patterns. First, both families peak near +100 synthetic samples (LLM: higher F1/recall; GAN: balanced F1/MCC). Larger scales (+200) degrade PR metrics, suggesting distributional noise. Second, LLM-based augmentation tilts toward recall/F1 (broader coverage), while GAN-based augmentation tilts toward precision (fewer alerts but more misses). The choice depends on whether the HR program prioritizes coverage or intervention efficiency.

Table 2. LLM- vs. GAN-based augmentation with AutoGluon. Metrics include ROC-AUC, Accuracy, Precision, Recall, F_1, Balanced Accuracy, and MCC. Results suggest an optimal region near $+100$ samples, with degradation at $+200$.

Method	Aug. Size	ROC AUC	Acc	Precision	Recall	F_1	MCC
Raw Dataset	0	0.903	0.900	0.844	0.409	0.551	0.544
	50	**0.906**	0.907	0.857	0.455	0.594	0.582

Method	Aug. Size	ROC AUC	Acc	Precision	Recall	F_1	MCC
	100	0.902	**0.914**	0.833	**0.530**	**0.648**	**0.622**
	200	0.897	0.900	0.824	0.424	0.560	0.546
	50	0.904	0.912	**0.966**	0.424	0.589	0.607
	100	0.901	0.912	0.865	0.485	0.621	0.607
	200	0.897	0.909	0.842	0.485	0.615	0.596

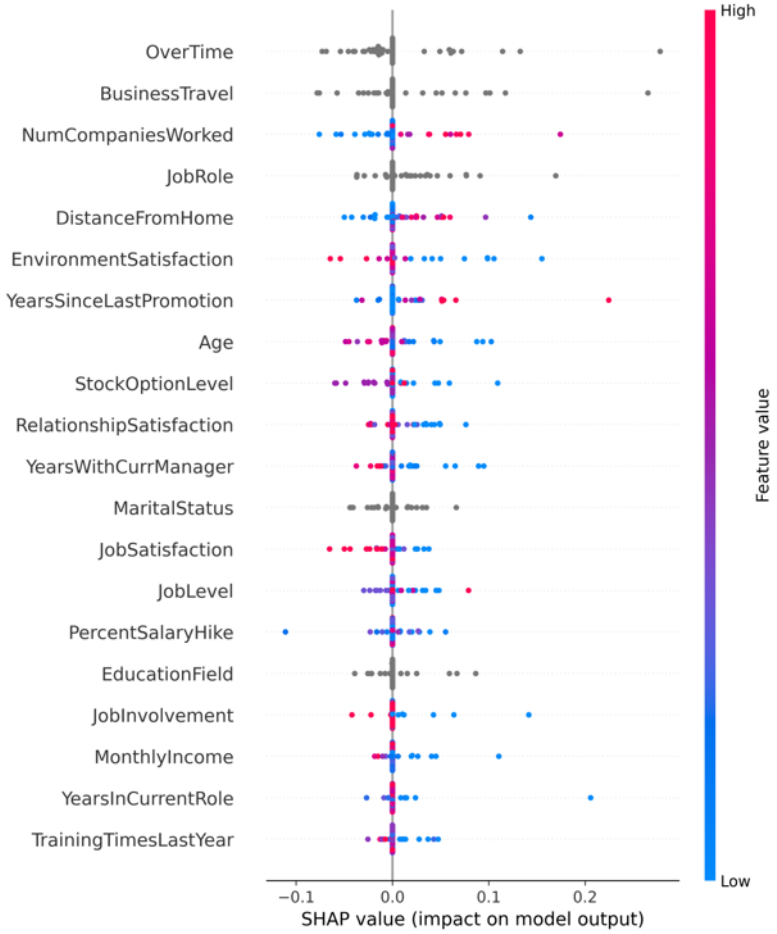


Figure 6. SHAP summary for AutoGluon-Tabular. Features are ordered by mean absolute SHAP. Red points indicate higher feature values; blue points indicate lower values. Key features such as OverTime, BusinessTravel, NumCompaniesWorked, JobRole, and DistanceFromHome dominate attributions

Table 2 summarizes these findings numerically. Both augmentation families peak near $+100$ synthetic samples (AutoGluon): the LLM-based approach attains its highest recall (0.530) and F_1 (0.648) at $+100$, whereas the GAN-based approach achieves higher precision (up to 0.966 at $+50$) with slightly lower recall. In practical terms, LLM-based augmentation broadens coverage of attrition cases (higher recall and F_1) at the expense of some precision, while GAN-based augmentation produces very precise alerts but misses more positives. For HR management, this means the LLM-driven method may flag more truly at-risk employees (potentially improving retention outreach) whereas the GAN-driven method yields fewer false alarms (potentially conserving intervention resources). The choice thus depends on whether the application prioritizes coverage (recall) or alert precision. Notably, adding too many synthetic cases ($+200$) degrades precision–recall performance for both methods, suggesting diminishing returns and the importance of quality control for generated data.

6.6.1 Augmentation on an Extended Dataset

To assess scalability, we repeated the augmentation experiments on an enlarged dataset (15k records). Table 3

reports performance for AutoGluon with and without augmented samples on this larger sample. Here the baseline task is already nearly solved (raw data AUC $= 0.9949$, accuracy $= 0.989$). Synthetic augmentation yields only marginal improvements: for instance, LLM-based augmentation at $+100$ raises accuracy from 0.9890 to 0.9903 and ROC-AUC from 0.9949 to 0.9955, with a slight F_1 increase from 0.976 to 0.979. GAN-based augmentation shows a similar pattern, and by $+200$ samples all methods saturate. These results suggest diminishing returns: once ample real data is available, adding synthetic examples yields very small gains in predictive power.

From an HR analytics perspective, the implications are two-fold. First, with a sufficiently large historical dataset, attrition prediction can be extremely accurate, reducing the marginal benefit of synthetic data. In practice, this means organizations with rich HR data may focus on feature engineering or model fine-tuning rather than heavy augmentation. Second, even on this larger scale, the slight improvements persist through $+100$ samples (LLM) before plateauing, implying that modest augmentation can still refine decision thresholds. Ultimately, however, the high accuracy across the board indicates that the model's insights (e.g., important predictors of attrition) are already robust in the large-data regime.

6.6.2 Effect of Critic-Filter on Augmentation

Finally, we evaluate the impact of our critic-filter, which removes implausible or duplicate synthetic records. Table 3 compares model performance when training on LLM-augmented data with and without this filter. The filtered case yields **higher recall** (0.530 vs 0.485) and **higher** F_1 (0.648 vs 0.634) than the no-filter case. Precision drops from 0.914 to 0.833 with filtering, reflecting the inclusion of more diverse (and some noisy) examples. Balanced accuracy improves from 0.738 to 0.756. The ROC AUC is nearly unchanged (0.9018 vs 0.8989), indicating that the overall ranking ability of the model is stable.

Table 3. Performance comparison between No-Filter LLM and Filter LLM (AutoGluon-Tabular). The critic filter increases recall and F_1 at the expense of precision.

Metric	**No-Filter LLM**	**Critic-Filter LLM**
Accuracy	0.9161	0.9138
F_1 Score	0.6337	0.6481
Precision	0.9143	0.8333
Recall	0.4848	0.5303
Balanced Accuracy	0.7384	0.7558
ROC AUC	0.8989	0.9018
MCC	0.6293	0.6218
ROC AUC (Curve)	0.8989	0.9018

These changes suggest that critic-filtering yields a model that flags a larger share of actual attrition cases (recall ↑) while generating more false positives (precision ↓). For HR practitioners, this means filtered augmentation helps catch more at-risk employees (which could improve retention reach), at the cost of additional interventions for those ultimately not leaving. The higher balanced accuracy and F_1 indicate better overall sensitivity-specificity trade-off. Organizations might choose filtering if missing an at-risk employee is considered costlier than investigating an extra false positive, especially since the filter's diverse synthetic cases make the model more robust to variant profiles.

6.5 Interpretability of Model Predictions

To ensure transparency, we examine SHAP for AutoGluon-Tabular (Figure 6). *OverTime*, *BusinessTravel*, *NumCompaniesWorked*, *JobRole*, and *DistanceFromHome* have the largest effects. Signs align with HR intuition: more overtime, frequent travel, many past employers, certain roles, and long commutes raise attrition risk; higher job satisfaction and tenure (not shown) reduce it. Sensitive attributes (e.g., gender) are not top drivers, which lowers fairness concerns and yields actionable levers for retention efforts.

6.6 GReaT vs. GAN: Practical Comparison (Continued)

GReaT fine-tuning is straightforward and stable, yielding diverse, plausible samples. CTGAN is powerful but requires careful tuning and can produce near-duplicates or constraint violations that need filtering. Multiple LLM sampling runs are more consistent than GAN runs in practice. For reuse, LLM-based augmentation adapts with

light prompting or brief fine-tuning; GANs typically need full retraining.

Before presenting the results of our LLM-based approach, we first provide a quantitative overview of its resource consumption. Compared with the traditional CTGAN method, our GReaT-based fine-tuning process required approximately 2 hours for training and generated about 100 samples per hour. The GPU memory usage during fine-tuning was around 38 GB, while data generation required about 20 GB. These metrics help illustrate the computational characteristics of our LLM-driven data generation pipeline.

Table 2 shows two patterns. First, both families peak near $+100$ synthetic samples (*LLM*: higher F_1/recall; *GAN*: balanced F_1/MCC). Larger scales ($+200$) degrade PR metrics, suggesting distributional noise. Second, LLM-based augmentation tilts toward recall/F_1 (broader coverage), while GAN-based augmentation tilts toward precision (fewer alerts but more misses). The choice depends on whether the HR program prioritizes coverage or intervention efficiency.

7. Discussion

Our findings show that LLM-based augmentation (GReaT) can meaningfully support decision-centric attrition prediction, with the most consistent benefit being an increase in recall (sensitivity) for the minority class. Across experiments, augmented models caught more true positives—i.e., more at-risk employees—than their non-augmented counterparts. In decision contexts where missed leavers (false negatives) cost more than false alarms, this gain in recall is practically valuable. For example, recall improvements from 60% to 70% among 100 potential leavers translate into preventing 10 extra departures, which can yield large savings relative to the incremental costs of false positives.

Selective Metric Gains. It is important to note that this recall improvement does not extend uniformly to all metrics. Discrimination measures such as ROC-AUC remained essentially unchanged or slightly lower with LLM augmentation, and Average Precision (AP) showed no significant uplift. Statistical tests (DeLong's test) confirm that observed AUC differences between augmented and non-augmented models are not statistically significant. This indicates that the main performance shift is selective: recall improves while AUC and AP remain stable. The underlying mechanism is intuitive: oversampling with synthetic positives shifts the classifier's boundary to capture more positives, raising sensitivity but at the cost of a modest increase in false positives. Precision therefore does not improve proportionally, explaining why composite metrics like AP and AUC stay flat.

Cost–Benefit Implications. From a managerial perspective, the asymmetry between false negatives and false positives provides the rationale for prioritizing recall. Industry evidence estimates replacement costs at roughly $0.5–2.0 \times$ annual salary and pegs U.S. annual voluntary turnover losses near $1T. (Note 2) If recall gains prevent 10 departures at $60k each (saving $600k), while five extra false positives each cost $5k in intervention ($25k), the net benefit is about $575k. This stylized example illustrates why recall-oriented augmentation strategies can deliver strong ROI despite modest precision trade-offs.

Model Interpretability and Feature Insights. SHAP analyses confirm that both augmented and baseline models identify similar, intuitive drivers (*OverTime*, *BusinessTravel*, *NumCompaniesWorked*, satisfaction and tenure signals). The consistency of feature importance across augmented and non-augmented runs supports face validity and stakeholder trust, ensuring that recall gains do not come at the expense of interpretability.

Comparison to Other Augmentation Methods. Relative to CTGAN, GReaT is operationally simpler (no adversarial dynamics), supports fully conditional sampling, and produced diverse yet plausible records. CTGAN remained competitive but required careful tuning and occasionally generated near-duplicates at small scales. Both methods show characteristic trade-offs: GAN-based augmentation leaned toward higher precision at small scales, while LLM-based augmentation favored recall/F_1. Over-augmentation hurt both, highlighting the need to *tune* augmentation rather than maximize it.

The Role of Critic-Guided Feedback. The Synergy of Actor-Critic Dynamics. Our results demonstrate that the raw output of autoregressive LLMs, while semantically coherent, is prone to 'over-exploration'—generating plausible but statistically unlikely outliers (hallucinations). The introduction of the Critic module effectively creates a feedback control mechanism during the augmentation process.

By leveraging the Critic's learned decision boundary, we enforce a strict quality gate that prioritizes distributional adherence over mere plausibility. This aligns with the principles of cybernetic control, where a system requires both an actuating signal (generation) and a correcting signal (rejection) to maintain stability.

- The Actor (LLM) contributes sensitivity by covering the 'long tail' of the minority class features that traditional GANs might miss (mode collapse).
- The Critic (Discriminator) contributes precision by rejecting samples that drift into the majority class space or undefined regions.

As shown in Table 4, this dual-network collaboration directly translates to a superior trade-off between Recall and

Precision. The Critic does not merely filter noise; it actively shapes the augmented decision boundary, ensuring that the synthetic samples used for training are those that provide the maximum information gain without introducing label noise.

Practical Limitations. LLM-based augmentation adds compute costs (fine-tuning and sampling) and requires disciplined validation (schema checks, deduplication) to avoid artifacts. Benefits are model- and dataset-dependent: deeper models like TabTransformer were more sensitive to heavy augmentation. We therefore recommend case-by-case evaluation, tuning augmentation size and weighting, and considering hybrid approaches (e.g., mixing in synthetic negatives).

Managerial Takeaways. When budgets are tight and false alarms must be minimized, precision-leaning augmentation (e.g., small-scale GAN-based) may be preferable. When the priority is preventing as many departures as possible, recall-leaning augmentation (moderate, quality-controlled LLM-based) is attractive. In all cases, pair augmentation with fairness monitoring and SHAP explanations so decisions remain equitable and actionable. Overall, repurposing LLMs for tabular augmentation offers HR analytics teams a pragmatic, recall-oriented tool—complementary to AutoML/ensembles—that merits deployment under leakage-aware protocols.

8. Conclusions

This study shows that *LLM-based augmentation* in the style of GReaT (Borisov, Seßler, Leemann, Pawelczyk, & Kasneci, 2023) can raise minority-class recall for employee attrition under severe imbalance. By adding realistic synthetic "leaver" records, augmented training flagged more at-risk employees, with *modest* precision trade-offs and generally preserved calibration. A leakage-aware protocol ensured that gains did not arise from target leakage, supporting the robustness of the findings.

Our contributions are fourfold. (i) We operationalize a *leakage-aware* modeling and evaluation framework that yields fair estimates on small, imbalanced tabular data. (ii) We assess fairness and interpretability—via equal-opportunity style group metrics (Hardt, Price, & Srebro, 2016) and SHAP explanations (Lundberg, & Lee, 2017)—to verify that augmented predictions remain equitable and transparent. (iii) We benchmark classical, boosted-tree, transformer, and AutoML ensembles, and show that LLM-augmented training improves minority recall and balanced accuracy over non-augmented baselines. (iv) We analyze augmentation scale and identify a practical operating region: recall gains plateau beyond moderate levels, and over-augmentation can degrade PR metrics. Augmentation should therefore be tuned rather than maximized. Together, these results integrate modern generative AI into an end-to-end, decision-oriented HR analytics pipeline.

From a decision-support perspective, higher recall means surfacing more would-be leavers for timely intervention—valuable in settings where false negatives are costlier than false positives. Notably, these gains were achieved while preserving interpretability (feature attributions aligned with domain knowledge via SHAP) and while keeping subgroup TPR gaps small (consistent with equal-opportunity fairness). Looking ahead, the same LLM-driven tabular augmentation can extend to other rare-event domains (e.g., churn, fraud, or risk) and can pair with prescriptive policies (thresholds and action rules) to move from prediction to intervention. Targeted conditioning during generation (e.g., subgroup-aware augmentation) and quality control (e.g., critic-based filtering) are promising next steps to further improve utility and fairness.

Overall, this work shows how cutting-edge generative AI can be *woven into* established analytics workflows to support managerial decisions. In line with DSAM's mission, our leakage-aware, interpretable, and fairness-conscious augmentation helps bridge advanced methods and practical HR deployment, offering a scalable path to more informed, equitable, and cost-effective retention strategies.

References

Azadi, S., Olsson, C., Darrell, T., Goodfellow, I., & Odena, A. (2018). Discriminator rejection sampling. *arXiv preprint arXiv:1810.06758*.

Barr, A. A., Rozman, R., & Guo, E. (2025). Generative adversarial networks vs large language models: A comparative study on synthetic tabular data generation. *arXiv preprint arXiv:2502.14523*.

Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M., & Kasneci, G. (2023). Language models are realistic tabular data generators. In *The eleventh international conference on learning representations,* ICLR 2023, kigali, rwanda, May 1-5, 2023, OpenReview.net.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357. https://doi.org/10.1613/jair.953

Chen, C., Liaw, A., Breiman, L., *et al.*. (2004). Using random forest to learn imbalanced data. *University of*

California, Berkeley, *110*(1-12), 24.

Chen, T., & Guestrin, C. (2016). A scalable tree boosting system. In *Proc 22nd ACM SIGKDD int conf knowl discov data min*, pp. 785-794. https://doi.org/10.1145/2939672.2939785

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd international conference on machine learning*, pp. 233-240. https://doi.org/10.1145/1143844.1143874

Demircioğlu, A. (2024). Applying oversampling before cross-validation will lead to high bias in radiomics. *Scientific Reports*, *14*(1), 11563. https://doi.org/10.1038/s41598-024-62585-z

Ding, X., Wang, Y., Wang, Z. J., & Welch, W. J. (2023). Efficient subsampling of realistic images from GANs conditional on a class or a continuous variable. *Neurocomputing*, *517,* 188-200. https://doi.org/10.1016/j.neucom.2022.10.070

Erickson, N., *et al.*. (2020). Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*.

Fabian, P. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research, 12*, 2825.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259-268. https://doi.org/10.1145/2783258.2783311

Flach, P. A., & Kull, M. (2015). Precision-recall-gain curves: PR analysis done right. In *Advances in neural information processing systems 28: Annual conference on neural information processing systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 838-846.

Glenn, W. B., *et al.*. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1), 1-3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, *29*.

Harter, J., & Adkins, A. (2019). This fixable problem costs U.S. Businesses $1 trillion. *Gallup Workplace*. Retrieved July 31, 2025, from https://www.gallup.com/workplace/247391/fixable-problem-costs-businesses-trillion.aspx

Hu, E. J., *et al.*. (2022). LoRA: Low-rank adaptation of large language models. In *The tenth international conference on learning representations,* ICLR 2022, virtual event, April 25-29, 2022. OpenReview.net.

Huang, X., Khetan, A., Cvitkovic, M., & Karnin, Z. (2020). Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*.

IBM Watson Analytics. (2015). IBM HR analytics employee attrition & performance dataset. Dataset. Retrieved July 31, 2025, from https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), p. 20150202. https://doi.org/10.1098/rsta.2015.0202

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Molodianovitch, K., Faraggi, D., & Reiser, B. (2006). Comparing the areas under two correlated ROC curves: Parametric and non-parametric approaches. *Biometrical Journal*, 48(5), 745-757. https://doi.org/10.1002/bimj.200610223

NeurIPS. (2025). NeurIPS paper checklist guidelines. Retrieved from https://neurips.cc/Conferences/2025/PaperInformation/PaperChecklist

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3), p. e0118432. https://doi.org/10.1371/journal.pone.0118432

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. *Advances in neural information processing systems*, *29.*

Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., & Flach, P. (2023). Classifier calibration: A survey on how to assess and improve predicted class probabilities. *Machine Learning*, *112*(9), 3211-3260. https://doi.org/10.1007/s10994-023-06336-7

Tang, Z., *et al.*. (2024). Autogluon-multimodal (automm): Supercharging multimodal automl with foundation

models. *arXiv preprint arXiv:2404.16233*.

Turner, R., Hung, J., Frank, E., Saatchi, Y., & Yosinski, J. (2019). Metropolis-hastings generative adversarial networks. In *International conference on machine learning*, PMLR, pp. 6345-6353.

van der Maaten, L., & Hinton, G. (2008, November). Visualizing data using t-SNE. *Journal of machine learning research*, *9*, 2579-2605.

XGBoost Developers. (2025). XGBoost parameters. Retrieved August 21, 2025, from https://xgboost.readthedocs.io/en/stable/parameter.html

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *Advances in neural information processing systems*, *32*.

YData. (2025). YData-synthetic: Synthesize tabular data. Retrieved August 21, 2025, from https://github.com/ydataai/ydata-synthetic

Notes

Note 1. At the GAN optimum $D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_g(x)}$, so $\frac{D^*(x)}{1-D^*(x)} = \frac{p_{\text{data}}(x)}{p_g(x)}$.

Note 2. Leaders and scarce roles tend to sit near the upper end.