P-TA: Using Proximal Policy Optimization to Enhance Tabular Data Augmentation via Large Language Models

Shuo Yang, Chenchen Yuan, Yao Rong, Felix Steinbauer and **Gjergji Kasneci**

Technical University of Munich, Germany

{shuo.yang, chenchen.yuan, yao.rong, felix.steinbauer, gjergji.kasneci}@tum.de

Abstract

A multitude of industries depend on accurate and reasonable tabular data augmentation for their business processes. Contemporary methodologies in generating tabular data revolve around utilizing Generative Adversarial Networks (GAN) or fine-tuning Large Language Models (LLM). However, GAN-based approaches are documented to produce samples with common-sense errors attributed to the absence of external knowledge. On the other hand, LLM-based methods exhibit a limited capacity to capture the disparities between synthesized and actual data distribution due to the absence of feedback from a discriminator during training. Furthermore, the decoding of LLM-based generation introduces gradient breakpoints, impeding the backpropagation of loss from a discriminator, thereby complicating the integration of these two approaches. To solve this challenge, we propose using proximal policy optimization (PPO) to apply GANs, guiding LLMs to enhance the probability distribution of tabular features. This approach enables the utilization of LLMs as generators for GANs in synthesizing tabular data. Our experiments demonstrate that PPO leads to an approximately 4% improvement in the accuracy of models trained on synthetically generated data over state-of-the-art across three real-world datasets.

1 Introduction

With the evolution of business processes that operate on structured data, tabular data has emerged as one of the most crucial data forms, owing to its easily manageable structure and efficient retrievability (Gilbert, 2022). Industries such as finance and healthcare heavily rely on tabular data for its clarity and ability to facilitate data analyses and comparisons (Shwartz-Ziv and Armon, 2022).

However, in realistic applications, the availability of high-quality tabular data is still an issue, mainly due to high costs of data collection and

annotation and privacy policies (Jensen and Potts, 2004). Therefore, various data augmentation methods have been proposed to overcome the scarcity of high-quality tabular data (Wen et al., 2022; Esmaeilpour et al., 2022).

The contemporary augmentation methods primarily revolve around rules, variational autoencoders (VAE), generative adversarial networks, and LLMs (Wei and Zou, 2019; Kingma and Welling, 2014; Patki et al., 2016; Bao et al., 2019; Borisov et al., 2023). However, these methods exhibit crucial limitations: (1) Rule-based methods rely on predefined constraints, which need manual effort and might not capture the diversity of real-world data (Wei and Zou, 2019). (2) The existing VAE-based and GAN-based approaches primarily focus on matching the distribution of table features rather than whether they are logically coherent (Kingma and Welling, 2014; Patki et al., 2016). For example, they might synthesize a sample with an "*Age*" of "*16*" and an "*Occupation*" of "*Professor*" (Goodfellow et al., 2014), resulting in implausible instances due to the lack of external knowledge (Bao et al., 2019). (3) Fine-tuning a LLM on real-world datasets may not capture the real-time differences in the distribution of features between synthetic and real-world data, thereby diminishing the quality of synthesized data. Furthermore, the generation of tabular features via LLMs is implemented through random sampling of the logits of decoders. However, sampling operations are non-differentiable, making it challenging to incorporate GAN-based methods for optimizing a generator, i.e., the LLM, based on the discriminator's output.

To overcome all of these drawbacks, we propose a method of training PPO-guided Language models for Table Augmentation (P-TA). Specifically, we first transform the tabular data into text using templates. After that, we fine-tune a LLM to generate new data in textual form. Additionally, we train a classifier to distinguish between actual and gener-

ated data, utilizing the outputs of the classifier as rewards for further optimizing the LLM through the PPO algorithm (Schulman et al., 2017).

Our methodology offers a distinct advantage: PPO-guided LMs can introduce external knowledge, and extensively explore various potential feature combination strategies during training by incorporating GANs, consequently enhancing the likelihood of producing high-quality synthetic data (Sutton et al., 1999).

As an additional study for reasonability, we propose to clarify the rationale behind tabular features to enhance the credibility of synthetic data by generating explanations. Explanations can serve as a data auditing tool, specifically for assessing the quality of synthesized data. The data auditing, in turn, contributes to enhancing the accountability of models trained using the synthesized data (Werder et al., 2022). We found the textual explanations are empirically more helpful for data auditors in assessing data quality than plain tables (see user study in our experiments). We are the first to integrate data auditing into tabular augmentation, distinguishing our work from previous efforts.

Our **contributions** are as follows:

- We propose P-TA, a novel framework to use PPO-guided language models for tabular augmentation, incorporating GANs to LLMs to generate tabular rows.
- We benchmark our framework with state-of-the-art (SOTA) methods on three realistic datasets. Training with our synthesized data improves the accuracy by 4% compared to the SOTA baseline, highlighting the high quality and potential impact of our approach in practical applications.
- We are the first to explain tabular data via retrieval enhancement. Our user study empirically shows that our explanations successfully assist users with auditing the generated data.

2 Related Work

With the development of databases, diverse sectors widely use tabular data as an indispensable data form (Gilbert, 2022). However, the accessibility of high-quality tabular data remains a persistent challenge, which is primarily attributed to privacy constraints and the substantial costs in collection (Jensen and Potts, 2004). These factors reflect the critical need to propose reasonable tabular synthesizing and auditing technologies.

2.1 Tabular augmentation

In prior work, table augmentation is mainly based on statistics (Kamthe et al., 2021), e.g., the Chow-Liu approximation (Chow and Liu, 1968). Contemporary approaches leverage heuristic algorithms, with methods based on VAEs (Ma et al., 2020; Darabi and Elor, 2021) or GANs (Choi et al., 2017; Park et al., 2018; Koivu et al., 2020). As representative approaches, TVAE (Xu et al., 2019) is optimized based on the evidence lower-bound (ELBO). It maps tabular features into a Gaussian distribution in the latent space and generates new data through sampling; CTGAN (Xu et al., 2019) adopts a GAN by conditioning the generation process solely on a single discrete feature for tabular data; CopulaGAN (Kamthe et al., 2021), the current state-of-the-art approach for GAN-based methods, simplifies the underlying CTGAN by utilizing Gaussian copulas. However, due to the lack of external knowledge, these models may generate tables with common-sense errors (Borisov et al., 2022). Consequently, we use LLMs as a knowledge base to introduce a PPO-guided table generator focusing on plausibility and common-sense knowledge. This makes the advantages of LLMs and GANs to be seamlessly integrated.

In the field of generative LLMs, prior to the integration of reinforcement learning, training methods primarily relied on masks (Kenton and Toutanova, 2019; Liu et al., 2019; Lewis et al., 2020) and autoregressive learning (Raffel et al., 2020; Brown et al., 2020). After that, the most advanced LLMs such as InstructGPT (Ouyang et al., 2022) and the family of LLaMA (Meta, 2023; Taori et al., 2023) demonstrated the powerful capabilities of the PPO algorithm mechanism.

2.2 Feature interpretation

Explainable Artificial Intelligence (XAI) methods have demonstrated their effectiveness in data auditing (Zhang et al., 2017). Shapley additive explanations (SHAP) Lundberg and Lee (2017); Nohara et al. (2019) quantify the impact of each feature on the model's output. Ribeiro et al. (2016) simulates complex models by training interpretable models to generate local explanations for specific data points. Gradient-based interpretability models, such as Grad-CAM (Selvaraju et al., 2017), IntegratedGrad (Sundararajan et al., 2017), generate heatmaps as explanations by computing gradients.
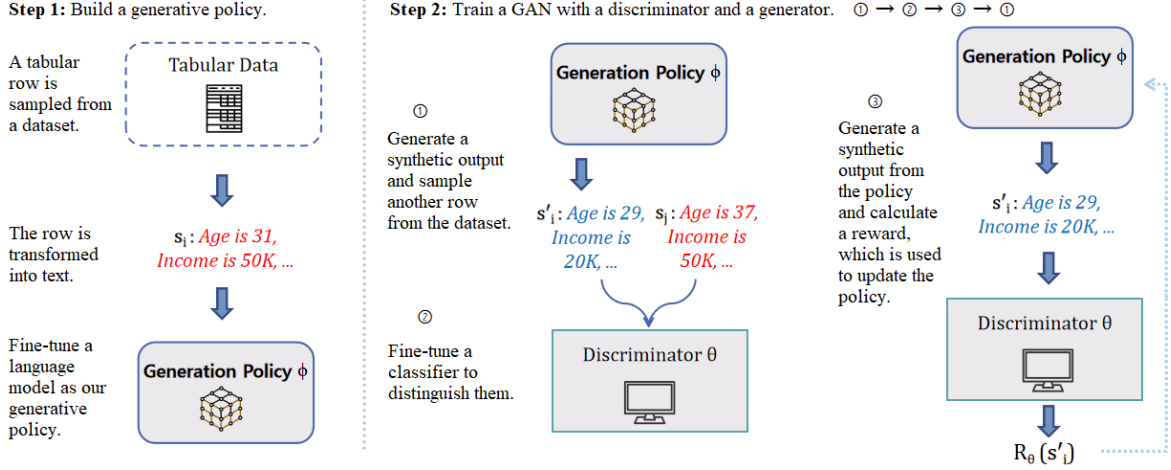
However, these methods emphasize feature at-

Figure 1: Training comprises two steps: 1) building an initial generative policy, 2) training a discriminator to distinguish the generated samples from ground truth, and updating the generative policy via PPO.

tribution concerning predictive impact but do not explain the meaningfulness or plausibility of features in a given context. Therefore, we elucidate the reasons behind inference results, aiming to enhance the credibility of our approach to generate textual explanations.

3 Methodology

Given a table with N rows and M columns, where each row represents a sample, and each column represents a feature. We denote this table as $T = (s_1, s_2, ..., s_N)$, where the i-th sample row is represented as $s_i = (a_{i1}, a_{i2}, ..., a_{iM})$, with a_{ij} denoting the j-th feature column of this sample.

Regarding data augmentation, our target is to generate a new table, denoted as $T' = (s'_1, s'_2, ..., s'_K)$, which contains K synthetic samples. The synthetic dataset should exhibit a distribution that resembles the original dataset T.

As for data auditing, our target is generating textual explanations for the plausibility of a given feature value a'_{ij} in an arbitrary new sample s'_i.

3.1 Tabular data augmentation

In the first step, we expand the tabular data by converting it into text through a predefined template. We then employ a LLM to generate new textual samples. Finally, we transform the generated text into tabular form via the same template. Figure 1 demonstrates the training pipeline.

Transforming tabular data into textual data. We employed an effective transformation template of "[Feature] is [Value]" (Borisov et al., 2023; Zhang et al., 2023). Here, [Feature] represents a specific feature name in a sample, and [Value] denotes the corresponding value. By connecting short

phrases associated with all features using commas, we construct a sentence describing a sample.

Training We initiate fine-tuning a LLM ϕ on these sentences. The LLM is treated as the generator of a GAN and a knowledge base to mitigate the logical and semantic inconsistencies or conflicts among the generated features (Heinzerling and Inui, 2021). Our objective is to enable the LLM to emulate the underlying distribution patterns of the feature values and minimize the generative perplexity, as shown in Eq. (1):

$$\mathcal{L}_{\text{LLM}}(\phi) = -\sum_i \log p(w_i^{\text{LLM}} = w_i | c_i; \phi), \quad (1)$$

where c_i is the context of a target word w_i. w_i^{LLM} is the corresponding prediction of the LLM.

We then generate new samples by sampling from the logits of LLM ϕ. The probability of generating the i-th token w_i as $w \in V$ is then given by $P_\phi(w|w_{1:i})$, where V represents the vocabulary. To encourage the LLM to generate diverse samples, we utilize a temperature function (Ackley et al., 1985) with a small τ:

$$P'_\phi(w_i) = \frac{P_\phi(w|w_{1:i})^{1/\tau}}{\sum_{w' \in V} P_\phi(w'|w_{1:i})^{1/\tau}}. \quad (2)$$

To prevent text degeneration, we employ the top-p sampling (Holtzman et al., 2020) and sample token w_i according to probability distribution P''_ϕ:

$$P''_\phi(w_i) = \begin{cases} P'_\phi(w_i) / \sum_{w' \in V_p} P'_\phi(w') & \text{if } w_i \in V_p \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

where top-p vocabulary V_p is the smallest set such that: $\sum_{w_i \in v_p} P(w_i|w_{1:i-1}) \geq p$.

Finally, we repeat the sampling and utilize the mentioned template to transform generated sentences back into tabular rows to obtain T'.

However, the missing or complex irregular dependencies between attributes in tables (Borisov et al., 2022) can lead to model biases and reduced robustness. Since GAN-based methods have proven to excel in generating text with given features (Yu et al., 2017), we employ the PPO algorithm to optimize the performance of the fine-tuned LLM (Ouyang et al., 2022) by incorporating it as the generator within a GAN.

In particular, we train a classifier θ to distinguish between synthetic data labeled as y^s and real data labeled as y^r. Here, θ is trained using a focal loss (Lin et al., 2020) to mitigate potential issues related to class imbalance. Then, we calculate the reward of a generated sample s' via the classifier θ:

$$\mathrm{R}_\theta(s') = P(y' = y^r | s'; \theta), \qquad (4)$$

where the reward value is the probability that y' is classified as actual data by the classifier θ.

We then employ PPO to increase the probability of generating sentences with high reward values, which theoretically leads to a closer alignment between generated and original data distributions. The objective function for model output s' consists of two components: the score $R_\theta(s')$ computed by the reward model and the KL divergence between the generative policy and the sampling policy, i.e., an orininal copy of the generative policy (Ouyang et al., 2022). Hence, the goal is to maximize:

$$E_{s' \in T'}[R_\theta(s') - \beta \log(P^{\mathrm{RL}}(s'; \phi)/P^{\mathrm{SFT}}(s'))], \qquad (5)$$

β is the weight coefficient, $P^{\mathrm{RL}}(s')$ and $P^{\mathrm{SFT}}(s')$ are the probability related to the generative policy and the sampling policy for s', respectively.

Here, we have two training objectives for the generator: (1) maximizing reward values by aligning distributions of synthetic data and actual data, and (2) minimizing the Kullback–Leibler divergence between the generation policy and the sampling policy.

Finally, we iterate through the training process until the GAN, i.e., the generator ϕ and the classifier θ, reaches a Nash equilibrium (Nash, 1950).

3.2 Data auditing

We aim to enhance the plausibility of our model by explaining and providing reasons for the generated feature values. On the one hand, reading these

---

Algorithm 1 Feature Interpretation Algorithm.

1: **Input:** A generated tabular row $s' = (a'_1, a'_2, ..., a'_M)$, with the j-th feature value a'_j to be explained; The original dataset $T = (s_1, s_2, ..., s_N)$.
2: **Output:** A textual explanation for why a_{ij} is reasonable in context s'_i.
3: # Generate textual descriptions.
4: Use prompt learning to transform s' and all $s_i \in T$ into their textual descriptions d' and $(d_1, d_2, ..., d_N)$.
5: # Retrieve descriptions using the augmented strategy.
6: Retrieve k most similar descriptions $(d_{\hat{1}}, ..., d_{\hat{k}})$ to d' from $(d_1, d_2, ..., d_N)$.
7: # Generate explanations based on descriptions.
8: Generate explanation using $(d_{\hat{1}}, ..., d_{\hat{k}})$ and d'.

---

explanations can enhance the transparency of table synthesis techniques for users. On the other hand, when this technology is applied to real-world data, it allows table users to assess the quality of tables without direct access to the table contents, thus ensuring data privacy in theory. We first describe each tabular row as a paragraph that rephrases all its features into text via a LLM. Then, the interpreter reads and compares these textual descriptions of the original samples to explain the reasons behind the presence of a particular value for a given feature. Algorithm 1 presents the steps towards such a plausibility-related interpretation.

Generation of textual descriptions. We first convert a table row into a sentence using the template described in Section 3.1, then fill it into a prompt template, e.g., "Please describe a person with the following features." Following our instructions, the interpreter pre-processes the templates into descriptions before generating the explanation.

Pre-processing tabular rows to descriptions for explanation holds three theoretical advantages: First, the LLM naturally incorporates additional information. For instance, for an individual with the job title "*professor*" age ranges between 0 and 25 years are less plausible, while a "*busy schedule*" may be quite likely. Therefore, these supplementary details can aid the interpreter in inferring more profound and intricate underlying reasons. Secondly, the tabular features always contain numerous abbreviations, many needing to be more readily understandable by language models and individuals who need domain expertise. By using a LLM

Travel Customers	https://www.kaggle.com/datasets/tejashvi14/tour-travels-customer-churn-prediction
Adult Income	https://archive.ics.uci.edu/dataset/2/adult
HELOC	https://kaggle.com/datasets/averkiyoliabev/home-equity-line-of-creditheloc

Table 1: The real-world datasets used in our study.

to convert them into text descriptions, they should be more comprehensible. Thirdly, (Dai and Callan, 2019) has demonstrated that using descriptive text outperforms using keywords (such as tabular feature values) in information retrieval tasks, which serves as empirical support for our discussion in the next paragraph.

Retrieval augmented generation. When explaining a generated feature value, we retrieve similar samples corresponding to a target sample as comparative data to our interpreter, an LLM. The advantage of retrieval lies in providing the interpreter model with comparative information, thereby preventing incorrect or incomplete analysis. For instance, if two individuals who are otherwise identical except for age receive different levels of "income", the interpreter can theoretically infer that age is the most significant factor causing the disparity in the "income" feature.

To retrieve similar samples, we compare two samples by comparing the semantic similarity of their descriptive texts. Specifically, we employ a LLM that has been fine-tuned on semantic similarity tasks to embed the descriptive texts corresponding to two samples separately. After that, we calculate the cosine similarity between the embeddings as a measure of sample similarity. Finally, the top k most similar samples retrieved are input into the interpreter for explanation.

4 Experiments

4.1 Datasets

We use three real-world datasets for evaluation by following (Borisov et al., 2023). The data we utilize can be directly downloaded from the internet, as shown in Table 1. The dataset **Travel** encompasses information about travel customers aimed at aiding a travel company in computing its churn rate. We utilize the "*Target*" feature, denoting customer attrition, as both the classification label and the feature to be audited. **Adult Income** consists of records from the 1994 Census database. It comprises 48,842 instances and 14 attributes. We treated the "*Income*" feature, denoting the annual income of specific individuals, as both the classifi-

cation label and the feature to be audited. **HELOC** dataset comprises 10,460 samples, each with 24 features related to anonymized information about *home equity line of credit* (HELOC) applications made by real homeowners. We regard the "RiskPerformance" feature as the classification label and the feature to be audited.

For evaluation, we generate 10,000 synthetic samples from Adult Income and HELOC and 1,000 synthetic samples from Travel Customers due to the relatively low data size and feature count in the latter dataset.

4.2 Automatic metrics

Accuracy. Following prior work (Borisov et al., 2023), we first conduct training using the original data on three classifiers: logistic regression (LR) (Cox, 1958), decision trees (DT) (Breiman et al., 1984), and random forests (RF) (Liaw et al., 2002). The training involve inputting the feature vectors of original samples into these classifiers and making predictions of the target feature. Subsequently, we supply the feature vectors of the generated samples as test samples to these classifiers and observe whether the predicted values by the classifiers match the generated values of the target variable. The closer the predictive accuracy remains to the original accuracy, i.e., accuracy on the original dataset, the more similar the distribution of the test dataset, i.e., the generated data, is to the distribution of the original dataset.

Discriminator measure. We use the accuracy of a classifier trained to distinguish between original and synthetic data to measure the performance of our generation approach. High-quality synthetic data should render the trained classifier incapable of accurately categorizing whether the data is synthetic or not.

To mitigate information leakage, we use a Kernel Support Vector Machine (VN, 2005), contrasting to the neural network-based classifier used in training.

Jaccard Coefficient. It is a statistical metric employed to quantify the similarity between two sets (Niwattanakul et al., 2013). We compute the Jaccard coefficient between the feature sets of a

Dataset		Original	TVAE	CopulaGAN	CTGAN	Neo + GReaT	GPT-2 + GReaT	Neo + Ours	GPT-2 + Ours
	LR ↑	84.29%	80.64%	74.23%	75.12%	74.82%	79.11%	<u>80.95%</u>	**82.22%**
Travel	DT ↑	86.91%	**80.90%**	75.61%	73.30%	72.73%	74.68%	74.81%	<u>80.14%</u>
	RF ↑	87.43%	<u>80.92%</u>	74.23%	72.23%	75.04%	75.31%	69.05%	**81.53%**
	Mean	86.21%	<u>81.01%</u>	74.69%	73.55%	74.19%	76.37%	74.94%	**81.30%**
	LR ↑	80.55%	<u>77.45%</u>	76.79%	**79.58%**	76.49%	75.69%	76.56%	76.27%
Adult Income	DT ↑	82.36%	<u>79.60%</u>	72.44%	78.50%	73.10%	73.17%	**79.67%**	73.27%
	RF ↑	85.81%	79.60%	77.46%	79.63%	<u>80.38%</u>	79.79%	**80.44%**	79.82%
	Mean	82.91%	78.88%	75.56%	**79.24%**	76.66%	76.22%	<u>78.89%</u>	76.45%
	LR ↑	69.79%	61.04%	42.03%	57.72%	<u>64.39%</u>	52.44%	**75.65%**	51.64%
HELOC	DT ↑	62.00%	**66.39%**	42.36%	61.34%	<u>60.16%</u>	56.51%	55.03%	57.14%
	RF ↑	70.12%	67.24%	42.35%	62.35%	<u>71.20%</u>	62.80%	**77.83%**	66.70%
	Mean	67.30%	64.89%	42.25%	60.47%	<u>65.25%</u>	57.25%	**69.50%**	58.49%

Table 2: Accuracy measure. LR, DT, and RF stand for Logistic Regression, Decision Tree, and Random Forest, respectively. **Bold** indicates the best performance, and <u>underline</u> indicates the second best in terms of being closest to the accuracy on the original data.

	CopulaGAN	CTGAN	TVAE	Neo + GReaT	GPT-2 + GReaT	Neo + Ours	GPT-2 + Ours
Travel ↓	97.58%	<u>92.13%</u>	**91.26%**	98.05%	98.04%	96.15%	96.26%
Adult Income ↓	88.47%	87.65%	96.92%	82.98%	<u>82.73%</u>	**82.51%**	<u>82.73%</u>
HELOC ↓	98.61%	100.00%	99.97%	92.79%	91.50%	**86.80%**	89.13%
Mean ↓	94.89%	93.26%	95.85%	91.27%	90.76%	**88.49%**	<u>89.37%</u>

Table 3: Discriminator measure. We use decreased accuracy values to check whether the data generated cannot be easily distinguished apart from the original data.

synthetic sample and the sample in the original dataset with the maximum feature overlap for that synthetic sample. A larger Jaccard coefficient indicates a closer similarity between the feature value distributions of synthetic data and the original data.

4.3 Human evaluation

We conduct data audits with human subjects by presenting them explanations generated by our model. Feedback from the pilot study indicates that auditing the HELOC dataset requires strong financial knowledge. Therefore, we conduct data audits on two datasets, Adult Income and Travel, for high-quality human evaluation. On each dataset, ten randomly selected samples are shown. For explanation quality assessment, human users rate the quality regarding **reasonableness, relevance, clarity, and comprehensiveness** within a 7-point Likert scale (Likert, 1932). Contrastively, we apply SHAP to generate numerical explanations. We asked all participants to indicate their preference for encountering either textual or quantified numerical explanations.

5 Results

5.1 Analysis for automatic evaluation

Table 2 illustrates the accuracy tests conducted on data generated by our approach, where our model outperforms baselines across three distinct datasets. Specifically, on the Adult Income dataset, our model achieves a considerable increase (of 4 percent points) in terms of accuracy, clearly surpassing the state-of-the-art methods. Similarly, our approach surpasses the baseline by 2% points on the HELOC dataset. In the case of the smaller-scale Travel dataset, our PPO techniques lead to a remarkable increase of 5% points. However, the performance of all the models on the HELOC dataset is lower. This should be attributed to the relatively high number of features in HELOC, not all correlated with the target feature being predicted. A substantial amount of redundant information from the multitude of features can potentially interfere with the predictions of these models.

In Table 3, it is observed that our method outperforms the baselines in terms of decreasing a discriminator's ability to discriminate between real and synthetic samples across all three datasets. Even in comparison to the best performing baseline (GPT-2 + GReaT), a reduction of over 2% points of the mean discrimination measure can be observed. In comparison with CopulaGAN or TVAE, the reduction is even more impressive (with over 6% and 7% points, respectively). This is an important advantage of our method, especially for practical applications, where a high reliability of synthesizers is needed. We attribute this to the close resemblance of the data generated by our approach to the original data distribution, rendering it challenging

Dataset	Ours					SHAP				
	REA.	COM.	CLA.	REL.	Mean	REA.	COM.	CLA.	REL.	Mean
Travel	4.29±0.19	4.32±0.19	4.87±0.18	4.37±0.20	4.46	3.69±0.21	3.93±0.21	3.11±0.24	4.26±0.22	3.75
Adult	4.26±0.20	4.00±0.22	5.10±0.17	4.22±0.20	4.40	3.64±0.20	4.17±0.21	3.14±0.23	3.74±0.20	3.67
Mean	**4.28**	**4.27**	**4.99**	4.19	**4.43**	3.67	3.84	3.13	**4.22**	3.71

Table 4: Human auditing results with 0.95 confidence interval, where "REA." represents reasonableness, "COM." represents comprehensive, "CLA." represents clarity, "REL." represents relevance.
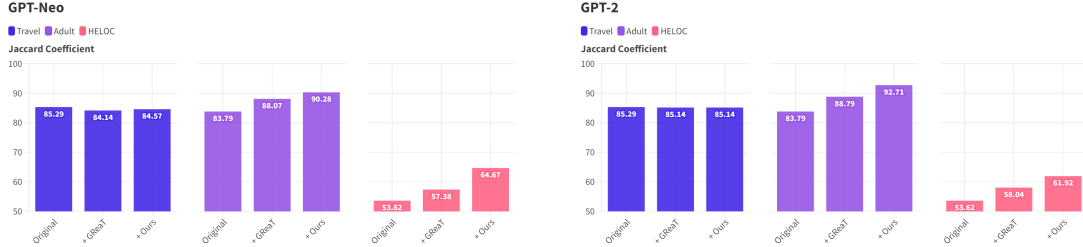

Figure 2: Jaccard Coefficient between generated and original samples in the training set. (Scaled by the factor of 100 for a clear view.)

for the discriminator to accurately distinguish between them. However, our method performs less effectively than TVAE on the Travel dataset. We hypothesize that when the number of tabular features is limited, TVAE exhibits similar or more effective data-fitting capabilities than LLMs.

5.2 Analysis for human evaluation

In the user study, we recruited 30 participants using online recruitment platform Prolific [1]. We required the participants to be fluent in English. Each participant was compensated with a payment of £6 for participation in the user study (within 40 minutes). As shown in Table 4, we observe that the explanations generated by our system receive positive evaluations (scores > 3.5) across all evaluation dimensions for both datasets. Compared to numerical SHAP explanations, our explanations achieve higher scores in clarity (CLA), reasonableness (REA) and comprehensiveness (COM), which contributes to a 19.4% improvement in the mean score. These results indicate the high quality of our explanations. Notably, 80% and 77.33% of individuals prefer our explanations for auditing the data to numerical explanations of SHAP in the Adult and Travel datasets, respectively.

When examining the REA score on each sample, there are samples that are not reasonable. For instance, in the Adult dataset, the example presented in Table 6 receives the lowest REA score. The synthetic data in this case indicates that corresponding adults have an income of "≥50K." However, all three binary classifiers trained for accuracy clas-

sify this example as "<50K," indicating that the example is an erroneous synthetic sample. Our users quickly identify logical errors by reading the explanation we provide.

	REA	COM	CLA	REL
SVM.	0.64	0.63	0.49	0.5

Table 7: Pearson correlation coefficient between human and automatic evaluation by an SVM model. All results yield p-value < 0.05.

To confirm that humans can effectively utilize our explanations to identify flaws in generated data that could impact classifier training, we calculate Pearson correlation coefficients among human evaluation metrics to the automatic evaluation by an SVM, which is the probability of a given sample being classified to that class. From Table. 7, we observe positive correlations, indicating that human and automatic evaluation of data quality exhibit consistency. This demonstrates that, by reading explanations generated by LLMs, people can assess the quality of tables without direct access to the contents, effectively preserving privacy of critical, person-related information. Notably, a robust correlation is found between the REA and ACC metrics, highlighting the power of our explanations.

5.3 Analysis for ablation experiment

In the augmentation task, as depicted in Fig. 2, we observe that PPO can further enhance the similarity of generated data to the original data distribution. This optimization benefits training on relatively large datasets, for instance, on the Adult and HE-

Travel	Synthetic data	age:49, workclass:Private, ... education:HS-grad, educational-num:9, marital-status:Married-civ-spouse, occupation:Craft-repair, ..., income:>50K
User	User's question	Explain the reason why the person has an income over 50K.
Ours	Explanation	... his educational background may have contributed to his ability to negotiate better pay and benefits packages ... given his marriage to someone with similar socioeconomic characteristics, such as being HS-grad and married ... increasing their joint income potential ...

Table 5: Data audit using our explanation on Adult Income Dataset. Blue text represents the main reason in the explanation. Please refer to appendix D for the full explanation.

Travel	Synthetic data	Age:30, ..., IncomeClass:Middle Income, ServicesOpted:2, AccountSynced-ToSocialMedia:No, BookedHotel:Yes
User	User's question	Explain the reason why the person hasn't churned.
Baseline	w/o. similar samples retrieved	... because they have booked a hotel, which indicates that they are still interested in the services offered ... their age falls within the middle income class ... It also does not appear that this customer is a frequent flyer or syncs their account to social media, both of which could potentially indicate higher levels of engagement with the company.
Ours	with similar samples retrieved	... because they have engaged positively by choosing two services ... which indicates some level of satisfaction ... despite not syncing their account with social media platforms, they have still managed to book a hotel ... indicating that the company's offerings meet their needs satisfactorily.

Table 6: Comparison of generated explanation on the Travel Dataset. Blue text represents key-points in explanations. Red text represents unreasonable explanations. Please refer to appendix D for more details.

LOC datasets, it improves the Jaccard similarity coefficient by 4% and 7%, respectively Additionally, we observe consistency in the performance of the two models in Fig. 2. Specifically, GPT-2 achieves higher performance on the Travel dataset, while GPT-Neo exhibites higher scores on the Adult and HELOC datasets. We hypothesize that this might be attributed to GPT-Neo's superior aptitude in handling longer tabular texts. The diversity in the training data for GPT-Neo may enhance its capability to capture textual features in longer documents.

Table 6 demonstrates the advantage of our proposed retrieval strategy. In this example, due to the absence of comparative information, the baseline confuses age and income in assessing the impact on user retention. Furthermore, we observe logical errors in explanations, such as treating "not a frequent flyer" as a positive factor. In contrast, our model analyzes all potential factors.

6 Limitation

During the training phase, as PPO involves the concurrent utilization of a discriminator and two policies, the time complexity for a single epoch is $2*O(k_1 n)+O(k_2 n)$, where k_1 and k_2 represent the time complexity of the generation policy and discriminator in training, respectively, and n denotes the number of features in the tabular data. In con-

trast, the time complexity for baselines generated based on language models is $O(k_1 n)$. In inference, our generation policy shares the same linear time complexity as the table generator GReaT, which is based on a language model. Utilizing a single A100 GPU, generating 1,000 HELOC tabular samples, each consisting of 24 features, incurs an average per-sample generation time of only 19.31 milliseconds. In summary, despite our model exhibiting higher training time complexity, it adequately meets the requirements of real-world applications.

7 Conclusion

Our research tackles critical challenges in tabular data usage, focusing on integrating language models into generative adversarial networks for data augmentation. Specifically, 1) Proximal Policy Optimization can leverage a discriminator's outputs to optimize a language model's generation strategy, thereby effectively leading to a substantial boost in the plausibility of generated data. 2) We successfully generated explanations for synthetic tabular features to assist data audits. Human evaluations highlight the clarity and the power of our explanations generated for data audits. These contributions could transform how industries obtain and audit tabular data, facilitating more informed decision-making and refined data-driven processes.

References

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169.

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xin-yu Dai, and Jiajun Chen. 2019. Generating sentences from disentangled syntactic and semantic spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6008–6019, Florence, Italy. Association for Computational Linguistics.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Rose Biderman. 2021. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2022. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21.

Vadim Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2023. Language models are realistic tabular data generators. In *The Eleventh International Conference on Learning Representations*.

L Breiman, JH Friedman, R Olshen, and CJ Stone. 1984. Classification and regression trees.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. 2017. Generating multi-label discrete patient records using generative adversarial networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 286–305. PMLR.

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

D. R. Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242.

Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for ir with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 985–988, New York, NY, USA. Association for Computing Machinery.

Sajad Darabi and Yotam Elor. 2021. Synthesising multimodal minority samples for tabular data. *ArXiv*, abs/2105.08204.

Mohammad Esmaeilpour, Nourhene Chaalia, Adel Abusitta, Franşois-Xavier Devailly, Wissem Maazoun, and Patrick Cardinal. 2022. Bi-discriminator gan for tabular data synthesis. *Pattern Recognition Letters*, 159:204–210.

Nigel Gilbert. 2022. *Analyzing tabular data: Loglinear and logistic models for social researchers*. Routledge.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1772–1791, Online. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Carlos Jensen and Colin Potts. 2004. Privacy policies as decision-making tools: An evaluation of online privacy notices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, page 471–478, New York, NY, USA. Association for Computing Machinery.

Sanket Kamthe, Samuel A. Assefa, and Marc Peter Deisenroth. 2021. Copula flows for synthetic data generation. *ArXiv*, abs/2101.00598.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Aki Koivu, Mikko Sairanen, Antti Airola, and Tapio Pahikkala. 2020. Synthetic minority oversampling of vital statistics data with generative adversarial networks. *Journal of the American Medical Informatics Association : JAMIA*, 27:1667 – 1674.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomforest. *R news*, 2(3):18–22.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Chao Ma, Sebastian Tschiatschek, Richard Turner, José Miguel Hernández-Lobato, and Cheng Zhang. 2020. Vaem: A deep generative model for heterogeneous mixed type data. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

AI Meta. 2023. Introducing llama: A foundational, 65-billion-parameter large language model. *Meta AI. https://ai. facebook. com/blog/large-language-model-llama-meta-ai*.

John F. Nash. 1950. Equilibrium points in <i>n</i>-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 380–384.

Yasunobu Nohara, Koutarou Matsumoto, Hidehisa Soejima, and Naoki Nakashima. 2019. Explanation of machine learning models using improved shapley additive explanation. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, page 546, New York, NY, USA. Association for Computing Machinery.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. 2018. Data synthesis based on generative adversarial networks. *Proc. VLDB Endow.*, 11(10):1071–1083.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. 2016. The synthetic data vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

Ravid Shwartz-Ziv and Amitai Armon. 2022. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vapnik VN. 2005. Universal learning technology: Support vector machines. *NEC Journal of Advanced Technology*, 2:137–144.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Bingyang Wen, Yupeng Cao, Fan Yang, Koduvayur Subbalakshmi, and Rajarathnam Chandramouli. 2022. Causal-TGAN: Modeling tabular data using causally-aware GAN. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*.

Karl Werder, Balasubramaniam Ramesh, and Rongen Zhang. 2022. Establishing data provenance for responsible artificial intelligence systems. *ACM Transactions on Management Information Systems (TMIS)*, 13(2):1–23.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. *Modeling Tabular Data Using Conditional GAN*. Curran Associates Inc., Red Hook, NY, USA.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Quanshi Zhang, Ruiming Cao, Feng Shi, Ying Nian Wu, and Song-Chun Zhu. 2017. Interpreting cnn knowledge via an explanatory graph. In *AAAI Conference on Artificial Intelligence*.

Tianping Zhang, Shaowen Wang, Shuicheng Yan, Jian Li, and Qian Liu. 2023. Generative table pre-training empowers models for tabular prediction. *arXiv preprint arXiv:2305.09696*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

A Ethics Statement

In the human evaluation phase, the individuals participating in the evaluation are anonymous to the authors of this paper. We pay them with the local standards and legal requirements. Furthermore, participants will only receive information about the data generated and the research questions mentioned above. We ensured that all participants comprehended the content of the task and agreed to participate in the manual evaluation.

Regarding model usage, we have obtained official approval from Meta to use LLaMA 2. Additionally, the Vicuna-1.5 model is an open-source supplementary model built upon the existing weights of LLaMA 2.

B Additional Experiments and Results

We employ the following metrics for implement additional experiments:

Average negative log-likelihood metric. The generated data is expected to conform to the distribution of the training data. Following (Borisov et al., 2023), we calculated the Log-likelihood of synthetic data samples on a density model derived from the original data (L_{syn}) and of the original test data on the model derived from the synthetic data (L_{test}) using a Gaussian Mixture Model for 10,000 samples generated on the Adult Income and HELOC datasets. As depicted in Table 8, our method demonstrates comparability with the state of the art (SOTA). Furthermore, we observed superior performance beyond the baseline concerning L_{syn}, although our approach exhibited slightly inferior performance compared to the baseline in

terms of L_{test}. Our analysis indicates that the PPO algorithm encounters challenges associated with overfitting in more prevalent data modes. Nevertheless, the adverse effects induced by PPO are relatively constrained, given its potential for a more pronounced enhancement in other utilized metrics.

Models	Adult		HELOC	
	L_{syn}	L_{test}	L_{syn}	L_{test}
Identity	-13.852	-13.852	-55.672	-55.672
CTGAN	-11.221	**-21.822**	-62.584	**-97.382**
GReaT-2	-10.706	-23.684	-61.600	-127.006
GReaT-Neo	-10.656	-26.054	-60.363	-134.411
Ours-2	-10.630	-24.095	-59.126	-135.449
Ours-Neo	**-10.484**	-28.433	**-58.260**	-147.703

Table 8: Average log-likelihood of synthetic and original data. We calculate them by using density models obtained from the original data (L_{syn}) and the synthetic data (L_{test}), respectively.

KL divergence. It quantifies the difference between two probability distributions (Kullback and Leibler, 1951). As most of the feature values in tabular data are not of numeric type and have relatively small value ranges, making visualization challenging, we select to visualize and calculate the KL divergence only for the distributions of "education years", "weekly working hours", and "age" in the Adult Income dataset.

Regarding consistency of our human evaluation, we computed Fleiss's Kappa for the scorings of reasonableness, relevance, clarity, and comprehensiveness, resulting in values of 0.16, 0.16, 0.21, and 0.15, respectively. Furthermore, for the Travel and Adult datasets, Fleiss's Kappa yielded values of 0.18 and 0.17, respectively. Since all the values are positive, we conclude that the participants exhibited consistency in their evaluations, thus affirming the reliability of human evaluation.

The Pearson correlation coefficient As a supplementary description to the Table 7, we present a heatmap illustrating the correlation coefficients between human evaluations and automated assessment metrics.

Repetition rate. It represents the proportion of duplicated instances in the generated samples compared to all generated samples. A well-performing system should exhibit a lower degree of replication.

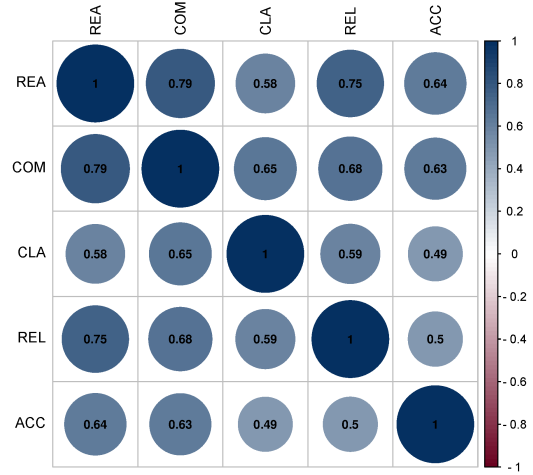**Area Under the Curve (AUC).** AUC is the area under the receiver operating characteristic curve.

Figure 3: The Pearson correlation coefficient between manual and automatic metrics. "ACC." is the probability of a given sample being classified as reasonable by an SVM. The p-value of significance is 0.05.

We use it to evaluate our model and the baseline to provide a comprehensive performance evaluation.

In terms of the performance of the binary classifier (less than 2%) trained on synthetic data, Table 9 illustrates that the data generated by our policy is similar to that of the baseline. However, it is worth noting that our model substantially improved the AUC metric of LR by over 30% on the Adult dataset, demonstrating the enhanced distributional consistency of our synthetic data, thereby enabling better applicability of the linear model. Furthermore, as the number of tabular features increased, the synthetic data generated by both methods tended to be more similar to the original data. It implies that language model-based table generation methods can exhibit advanced simulating capabilities when features are sufficient.

Considering the consistency of label distributions between synthetic and actual data, we first computed the KL divergence between the label distributions of synthetic data for testing and the original dataset. For the Adult dataset, the mean KL divergence was 0.01 ± 0.02 and a maximum value of 0.04. For the Travel dataset, the mean KL divergence was 0.02 ± 0.02 and a maximum value of 0.03. For the HELOC dataset, the mean KL divergence was 0.01 ± 0.01 and a maximum value of 0.03. Since the KL divergences between the label distributions of all synthetic datasets and the original dataset were less than 0.05, we conclude that the label distribution of the testing dataset is similar to that of the actual data.

The results presented in Table 10 demonstrate

Dataset	Travel				Adult Income				HELOC			
Model	LR ↑	DT ↑	RF ↑	Mean ↑	LR ↑	DT ↑	RF ↑	Mean ↑	LR ↑	DT ↑	RF ↑	Mean ↑
Original	0.85	0.87	0.95	0.89	0.58	0.76	0.91	0.75	0.76	0.62	0.78	0.72
GPT-Neo + GReaT	**0.76**	0.68	**0.72**	**0.72**	0.54	0.70	0.86	0.70	**0.66**	**0.60**	**0.78**	**0.68**
GPT-2 + GReaT	<u>0.69</u>	0.65	<u>0.68</u>	0.67	0.54	0.70	0.86	0.70	0.54	0.57	0.68	0.60
GPT-Neo + Ours	0.61	**0.79**	0.66	<u>0.69</u>	0.87	0.71	0.87	0.82	<u>0.64</u>	0.57	<u>0.77</u>	<u>0.66</u>
GPT-2 + Ours	0.62	<u>0.73</u>	0.63	0.66	0.57	0.68	0.82	0.69	0.54	0.57	0.72	0.61

Table 9: AUC of classifiers trained with synthesized data using different models (listed on the left) on three datasets.

Datasets	Travel	Adult Income	HELOC
Original	53.14%	0.11%	5.61%
Neo + GReaT	16.00%	0.01%	0.00%
GPT-2 + GReaT	22.25%	0.00%	0.00%
GPT-Neo + Ours	16.00%	0.01%	0.00%
GPT-2 + Ours	22.67%	0.00%	0.00%

Table 10: Repetition rate (with top-p of 0.9 and temperature of 0.9 in decoding).

that PPO does not significantly increase the replication rate (less than 1%). We observed that for Adult Income and HELOC datasets, our method seldom generates results identical to those in the original dataset. However, for the Travel dataset, given its mere seven features, with four having a binary domain, replication in the generation process unavoidably occurs.

C Implementation Details

Device	Server A	Server B
Memory	13G	1012G
CPU	Intel Core i5-1135G7 2.40G Hz	AMD EPYC 7763 2.45G Hz
GPU	NVIDIA Tesla T4 16G	NVIDIA A100 80G
Operating System	Ubuntu 18.04.6 LTS	Ubuntu 20.04.6 LTS

Table 12: Experimental hardware environment.

For the experimental setup, we utilized two servers. All work related to the significant language model mentioned in our paper was conducted on Server B. Details of the experimental environment are provided in Table 12.

We employed the standard GPT-2 (Radford et al., 2019) with 1.5 billion parameters and GPT-Neo (Black et al., 2021) with 1.3 billion parameters models as generators and a self-attention-based Convolutional Neural Network (CNN) (Lecun et al., 1998). The CNN has 32 convolutional kernels, eight attention heads, and 512 neurons in the attention layer (Vaswani et al., 2017).

For the interpretability task, we randomly selected 100 samples from the test set for explanation and human evaluation. The LLM and interpreter used for describing samples were Vicuna-v1.5 (Zheng et al., 2023), fine-tuned on LLaMA 2 (Meta, 2023). The model used for computing semantic similarity between descriptive texts was Sentence-Bert (Reimers and Gurevych, 2019).

During training, we utilized the optimizer of AdamW (Loshchilov and Hutter, 2018) with a learning rate of 1e-4 and a batch size of 16. During inference, the maximum generation length was 300, top-p (Holtzman et al., 2020) was 0.9, temperature was 0.7, and repetition penalty coefficient was 1.2.

The set of prompts we used for explanations is presented in Table 11.

Regarding the questionnaires used, we posed the following four statements for each sample evaluated:

(reasonableness) *The model explanation exhibits minimal logical errors and represents rational interpretations of the question.*

(relevance) *The model explanation does not include irrelevant content to the topic.*

(clarity) *The model explanation is easy for me to understand.*

(comprehensiveness) *The model explanation covers all important features as I would expect.*

To ensure the effectiveness of the feedback collected, we established a gold question and provided two options:

(gold question) *Before getting started, let's make sure that you understand the task. Please choose the task that you will do:*

A. *I will rate the quality of the decision provided by the model.*

B. *I will rate the quality of the explanation provided by the model.*

We will only consider feedback from respondents who answered the gold question correctly

with option A. Any feedback from respondents who omitted answers will also be deemed invalid.

Furthermore, we randomly selected a sample from the dataset, placing it at the beginning of the questionnaire as a warm-up example to ensure the respondents understood the task.

D Additional Case Study

To specifically illustrate the effectiveness of our explanations, we conducted a case study by randomly selecting samples from the three datasets.

The explanations generated are presented in Table 6, Table 13 and Table 14. We observed that our model precisely captured the underlying reasons that could lead to specific features and provided further implications based on these reasons. For instance, the model realized that booking a hotel room is a positive signal for customer retention. At the same time, low activity on social media may be a negative signal, while other features, such as age, may not significantly impact the retention of the target person. Additionally, through comparisons with other samples, the model recognized the significance of specific feature values. For instance, it noted that working 35 hours per week falls under the category of long working hours, and having an educational duration of 13 years is higher than the average level.

Travel Customers	Adult Income	HELOC
Your task is to Core Part, referring to the following set of 3 positive customers who haven't churned from a travel company and 3 negative customers who have churned from a travel company:	Your task is to Core Part, referring to the following set of 3 positive adults who earn annual incomes which exceed $50K and 3 negative adults who earn annual incomes which don't exceed $50K:	Your task is to Core Part, referring to the following set of 3 positive individuals who have never been late for payments by more than 90 days over a period of 24 months since the account of Home Equity Line of Credit (HELOC) was opened and 3 negative individuals who have been late for payments at least 90 days by at least once over a period of 24 months since the account of Home Equity Line of Credit (HELOC) was opened:
---	---	---
POSITIVE CUSTOMERS [1] FEATURES: "Features of customer 1" DESCRIPTION: "Description of customer 1" ... NEGATIVE CUSTOMERS [1] FEATURES: "Features of customer 4" DESCRIPTION: "Description of customer 4" ...	POSITIVE ADULTS [1] FEATURES: "Features of adult 1" DESCRIPTION: "Description of adult 1" ... NEGATIVE ADULTS [1] FEATURES: "Features of adult 4" DESCRIPTION: "Description of adult 4" ...	POSITIVE INDIVIDUALS [1] FEATURES: "Features of individual 1" DESCRIPTION: "Description of individual 1" ... NEGATIVE INDIVIDUALS [1] FEATURES: "Features of individual 4" DESCRIPTION: "Description of individual 4" ...
---	---	---
Your task is to Core Part.	Your task is to Core Part.	Your task is to Core Part.
Core Part: explain the reason why the customer with the FEATURES: "Features" and DESCRIPTION: "Description" has (hasn't) churned from the travel company	Core Part: explain the reason why the adult with the FEATURES: "Features" and DESCRIPTION: "Description" earns an annual income which doesn't exceed (exceeds) $50K	Core Part: explain the reason why the individual with the FEATURES: "Features" and DESCRIPTION: "Description" has been late for payments at least 90 days by at least once (never been late for payments by more than 90 days) over a period of 24 months since the account of Home Equity Line of Credit (HELOC) was opened

Table 11: Prompts used to generate explanations. Red text represents the core part of a prompt; brown text represents the features or description of a specific sample; orange text represents an alternate text determined by the target value of a sample.

Adult	Synthetic data	age:54, workclass:Local-gov, fnlwgt:176977, education:Bachelors, educational-num:13, marital-status:Married-civ-spouse, occupation:Prof-specialty, relationship:Husband, race:White, gender:Male, capital-gain:0, capital-loss:0, hours-per-week:35, native-country:United-States, income:>50K
	Description generated	John Doe is a 54 year old male from the United States who works as a professional specialist for local government. He has a Bachelor's degree and 13 years of formal education under his belt. John is married to his civilian spouse and is white in terms of both race and ethnicity. His weekly working hours are 35, with an annual income exceeding $50k.
User	User's question	Explain the reason why the adult has an income over $50K.
Ours	Explanation generated	The reason he earns an annual income which exceeds $50K is because he holds a position as a professional specialist for local government, where his expertise and experience likely command higher wages compared to other occupations within the same work class. Additionally, his completion of a Bachelor's degree and 13 years of formal education contribute to his knowledge base and skillset, making him more valuable to his employer. Furthermore, John's marriage to someone outside of the military may provide additional support and resources, helping him maintain a comfortable standard of living despite his relatively long working hours each week (35). Lastly, his status as a white male in the United States could be advantageous when negotiating salaries and benefits. Overall, these factors combine to give John Doe access to better paying jobs than those without similar qualifications, resulting in an annual income above $50K.

Table 13: Case Study of the Adult Income Dataset. Blue text represents key-points in explanations.

HELOC	Synthetic data	RiskPerformance:Good, ExternalRiskEstimate:69, MSinceOldestTradeOpen:184, MSinceMostRecentTradeOpen:5, AverageMInFile:74, NumSatisfactoryTrades:31, NumTrades60Ever2DerogPubRec:0, NumTrades90Ever2DerogPubRec:0, PercentTradesNeverDelq:100, MSinceMostRecentDelq:-7, MaxDelq2PublicRecLast12M:7, MaxDelqEver:8, NumTotalTrades:31, NumTradesOpeninLast12M:0, PercentInstallTrades:23, MSinceMostRecentInqexcl7days:-7, NumInqLast6M:0, NumInqLast6Mexcl7days:0, NetFractionRevolvingBurden:3, NetFractionInstallBurden:-8, NumRevolvingTradesWBalance:1, NumInstallTradesWBalance:-8, NumBank2NatlTradesWHighUtilization:0, PercentTradesWBalance:33
	Description generated	With an external risk estimate of 69 and a minimum file score of 74, it appears that this HELOC may pose some level of risk but is still within acceptable parameters. Additionally, there have been no trades delinquent by more than 60 days or those ever reported as public records in the last two years. Furthermore, only one trade is considered ʿderogatory,ʾ indicating that the borrower's credit history is generally positive. It's worth noting that this HELOC has had zero trades open within the past year, which could suggest a low utilization rate. However, the maximum delinquency recorded over the past twelve months was seven, suggesting potential room for improvement.
User	User's question	Explain the reason why the applicant will not have a 90-day or more severe delinquency after opening a credit account.
Ours	Explanation generated	The explanation lies in the fact that the individual with the given features has shown consistent and reliable repayment behavior throughout their credit history. Despite being delinquent in the past, they have since made up for it by demonstrating perfect payment history in the subsequent periods. Moreover, their current credit score indicates that they are likely to continue making timely payments going forward. As a result, despite the presence of certain risk factors like high revolving burden and negative install burden, their overall creditworthiness is deemed sufficient to justify the assigned risk estimate of 69.

Table 14: Case Study of the HELOC Dataset. Blue text represents key-points in explanations.