Exploring the Generation of Synthetic Educational Tabular Data using LLMs

Qinyi Liu
Centre for the Science of Learning & Technology (SLATE)
University of Bergen
Bergen, Norway
qinyi.liu@uib.no

Mohammad Khalil
Centre for the Science of Learning & Technology (SLATE)
University of Bergen
Bergen, Norway
mohammad.khalil@uib.no

ABSTRACT

Synthetic data as a means of data generation and augmentation has garnered widespread attention from various fields including educational data science. However, discussions about using large language models (LLMs) to generate synthetic data have primarily focused on textual data rather than tabular data. In educational data analysis, tabular data, commonly used by learning management systems, plays a crucial role. Therefore, exploring how to utilise advanced technologies of LLMs to generate high-quality synthetic tabular data is of great significance for advancing the field of educational data analysis. This paper compares the performance of an advanced LLM named GReaT and a key representative synthetic tabular data generation algorithm called CTGAN in generating educational tabular data. The aim is to explore the potential of the educational tabular data generated by these two models in terms of data resemblance, privacy, and predictive utility. This study fills a current research gap and practice in the application of LLMs in generating educational tabular data.

CCS CONCEPTS

• Applied computing~Education

KEYWORDS

Generative AI, Synthetic data, Large language Model (LLM), education, AI in Education (AIED)

ACM Reference format:

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK'18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

1 Introduction

Tabular data is critical across various fields, including education [7, 12], forming the backbone of numerous classification and regression machine learning applications for education data science. The superior performance of machine learning models is often built on the foundation of data, with the quality and quantity of data being the key drivers of their efficacy. However, in practical applications, obtaining sufficient and high-quality tabular data often faces multiple challenges, including high costs, time-consuming processes, and strict data privacy protection requirements [15]. These factors, either individually or collectively, lead to a common phenomenon in real-world scenarios: scarce feature information and imbalanced class distributions. This incompleteness and imbalance in data impact the training of machine learning models, making it difficult to achieve their optimal performance [12].

Synthetic data generation has emerged as a promising solution, offering a cost-effective way to enrich datasets [12]. For tabular data, traditional synthetic data generation methods fall into two categories: statistical model-based approaches and deep learning-based approaches (e.g., generative adversarial networks-GANs). Both approaches have shown promising progress in generating tabular datasets [10]. Large Language Models (LLMs) have emerged as one of the most promising approaches in deep learning. Nevertheless, LLMs have not yet demonstrated excellence in producing high-quality synthetic tabular data, particularly in terms of utility and fidelity. Recently, [2] have shown promising advancements in this area, challenging previous beliefs. The performance of Borisov's model for education still requires exploration.

Therefore, we select an advanced LLM that has shown promising performance in tabular data generation, called, GReaT developed by [2]. To validate GReaT, we also select a popular representative synthetic tabular generation algorithm (conditional tabular GAN short for CTGAN) [22] to generate data on educational datasets and conduct a comprehensive evaluation covering resemblance, utility, and privacy. Through this evaluation, we demonstrate that the used LLM (i.e., GReaT) performs well on educational tabular datasets, achieving performance in terms of resemblance, utility, and privacy that is comparable to, or even better than, some of the best-known tabular data generation algorithms. This study highlights a promising start for the application of LLMs in the research domain of educational tabular data. Our contribution can be summarised as the following:

1. Our research fills the gap in LLMs' application to tabular data generation in education. We show a case that advances educational data processing techniques and provides

examples of experiments using the latest data augmentation tools to drive the scaling of educational data analytics.
2. Our findings highlight that GReaT and CTGAN achieve comparable performance across multiple metrics, offering promising insights for the future application of LLMs in educational data.

2 Related work

2.1 Synthetic tabular data generation

Deep learning methods are one of the most important approaches for synthesising tabular data [10]. Within deep learning methods, GAN-based approaches are very popular and have a strong track record in both utility and privacy protection [1, 10, 18]. For tabular data generation, CTGAN stands out among GAN-based methods, outperforming other tabular synthesis algorithms such as TGAN and Tabular VAE [22]. CTGAN showcases unique features, including the ability to handle both discrete and continuous features, the use of a conditional generator, and sampling-specific training to avoid mode collapse and data imbalance issues [23]. Moreover, the synthetic data generated by CTGAN can be applied to various applications, such as data augmentation, data privacy, and data analysis [9]. CTGAN excels in preserving the underlying structure of real data, having been validated across multiple fields, sample sizes, and data balances [11, 23]. The CTGAN mathematical formulation is as follows, including the objective function that the GAN aims to optimise during the training process, which includes both discriminator and generator networks [22]. The discriminator network aims to distinguish between real and synthetic data, while the generator network aims to produce synthetic samples that can fool the discriminator.

$$\max_D \mathcal{L}_D = \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}\left[\log D(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{x}\sim G(\boldsymbol{z}),\boldsymbol{z}\sim p(\boldsymbol{z})}\left[\log(1 - D(\boldsymbol{x}))\right] \quad (1)$$

where G and D denote the generator and discriminator networks, respectively, and *P_data* and *P_z* are the distributions of real and synthetic data, respectively. The objective function is iteratively optimised until the generator produces samples that are indistinguishable from real data by the discriminator.

LLMs are defined as large models, typically transformers, trained on vast amounts of textual data to predict the next word given the preceding text [21]. Researchers have traditionally considered LLMs to be more adept at generating textual data, with less exploration into their performance in generating tabular data [4]. However, the remarkable performance of LLMs has recently garnered significant interest from both academia and industry, leading to the belief that LLMs might lay the groundwork for achieving Artificial General Intelligence [3]. Consequently, scholars have begun exploring the capabilities of LLMs in handling various tabular data tasks [4]. A notable example is the Generation of Realistic Tabular Data (GReaT), a framework proposed by [2] specifically designed for generating tabular data using LLMs. GReaT leverages an auto-regressive generative LLM to produce highly realistic tabular data. The advantages of GReaT, as demonstrated in experimental results, are significant. First, GReaT can generate data conditionally based on any subset of features, with the remaining features being sampled without additional computational overhead, making it highly effective in data imputation and oversampling tasks [2]. Second, due to its flexible feature-conditioned generation capability, GReaT can be applied to various data generation tasks, including handling imbalanced datasets, imputing missing data, and generating data with specific feature combinations. In Borisov's paper, GReaT has shown results in various datasets surpassing those of CTGAN. However, given the relative newness of this approach, it has not been tested on datasets from other domains (e.g. education), highlighting a gap that this paper aims to address.

2.2 Synthetic tabular data evaluation

The evaluation of synthetic data is generally conducted from three perspectives: resemblance (also referred to as fidelity or similarity), utility, and privacy [10]. Resemblance, as the name suggests, aims for the synthetic data to closely match the distribution structure of the real data. Utility means that the synthetic data should perform similarly to real data in data analysis tasks. In a broader sense, utility also encompasses resemblance as Jordon and colleagues explain [10]. Privacy ensures that the synthetic data, while closely approximating the real data distribution, does not leak sensitive personal information from the real data and remains robust against attacks. In this paper, to comprehensively compare the performance of the LLM-based GReaT model with that of CTGAN, we will select representative metrics from these three dimensions: resemblance, utility, and privacy for the evaluation. Details of the specific metrics and the reasons for their selection can be found in section 3.2.

3 Experiments

The experiments in this study utilised a Python library called Synthcity, developed by [19], which is dedicated to the convenient generation and evaluation of synthetic tabular data. Synthcity provides a comprehensive set of tools for creating high-quality synthetic datasets and includes features for assessing the resemblance, utility, and privacy of the generated data. The experiments were conducted in a Google Colab environment, leveraging its powerful computational resources. The GPU used for this study was the NVIDIA A100, which is known for its exceptional performance in handling large-scale machine learning and deep learning tasks. This setup ensured efficient execution of the experiments, allowing for rapid prototyping and extensive evaluations.

Regarding to the synthetic data generators used in the paper, as introduced in the 2.1, this paper utilizes two open-source synthetic data generators, GReaT and CTGAN. Both employing their recommended default parameters. In the GReaT framework, the LLM used is distilgpt2.

3.1 Dataset

The dataset used in this paper is the student performance dataset of CC-BY license that is publicly accessible on Kaggle [13]. The dataset includes information on parental background, exam preparation, and student performance. These variables are common in many educational tabular datasets. The specific information of this dataset is shown in Table 1.

Table 1: Details of the used dataset

Dataset variable	Details
Year	2018
Number of attributes	8
Number of records	1000
Target variables	continuous
Number of continuous variables	3
Number of categorical variables	5
Imbalance ratio	0.05

3.2 Evaluation methods

3.2.1 Resemblance

Average Jensen-Shannon Distance (JSD). JSD measures the similarity between two probability distributions [17]. It is a symmetrized and smoothed version of the Kullback-Leibler divergence. In tabular synthetic data, a lower JSD indicates that the synthetic data distribution closely approximates the real data distribution, ensuring higher resemblance. Additionally, JSD is particularly useful for addressing categorical data [10]. The mathematical definition of is given below [17]:

$$js\_dist(p,q) = \sqrt{\frac{D(p||m) + D(q||m)}{2}} \quad (2)$$

m is the pointwise mean of *p* and *q*, and D is the Kullback-Leibler divergence, defined in Equation 3. The probability distributions of the features have been used to compute these distances: *p* is the probability distribution of the real data attribute, and *q* is the probability distribution of the synthetic data attribute [6].

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \quad (3)$$

Maximum Mean Discrepancy (MMD). MMD is a distance measure between the sample means of two distributions in a reproducing kernel Hilbert space (RKHS) [5]. In tabular synthetic data, a lower MMD indicates that the generated synthetic data matches the overall statistical properties of the real data, ensuring the preservation of the data's global distribution [18]. MMD is versatile and can address both categorical and continuous data. The mathematical definition of MMD is given below [5]:

$$\mathrm{MMD}_k(P,Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}} . \quad (4)$$

P represents the distribution of the real data in this context, and *Q* represents the distribution of the synthetic data. The expression ‖ μP – μQ ‖ H measures the distance between the mean embeddings of *P* and *Q* in the RKHS.

Wasserstein Distance (WD). WD measures the minimum cost of transforming one probability distribution into another, also known as the Earth Mover's Distance [14]. In tabular synthetic data, a lower WD signifies that the synthetic data preserves the structure of the real data, enhancing the data's usability and integrity. WD is particularly useful for addressing continuous data to maintain similarity in distribution shape [24]. The mathematical definition is given below [14]:

$$was\_dist(r,s) = \int_{-\infty}^{+\infty} |R - S| \quad (5)$$

WD can be seen as the minimum cost required to transform a vector (*r*) into another vector (*s*), where the cost is measured as the amount of distribution weight that must be moved, multiplied by the distance it has to be moved (Equation 5); R and S are the cumulative distribution function of the real data, and synthetic data attributes, respectively [6].

3.2.2 Utility

One of the best methods for validating synthetic data utility is by evaluating its performance on a downstream Machine Learning (ML) task [9]. A common and classic approach for this is the Train-Synthetic-Test-Real (TSTR) evaluation proposed by [8]. The TSTR method involves five steps: First, a real dataset is split into a main dataset for training and a holdout dataset for evaluation. Next, a synthetic dataset is created based solely on the training data. Then, an ML model is trained—once using the synthetic data and once using the actual training data. The performance of each model is then evaluated against the actual holdout data. By comparing the performance of these two models, the utility retained by the synthesization method with respect to a specific ML task can be assessed [8]. In this paper, we will use frequently used machine learning algorithms from educational data mining and learning analytics to assess the utility of synthetic data, specifically Logistic Regression, Multi-Layer Perceptron (MLP), and Extreme Gradient Boosting (XGBoost).

3.2.2 Privacy

Membership Inference Attack (MIA) with Prior Knowledge. MIA evaluates the ability of an adversary to determine whether a particular data point was included in the training dataset used to generate the synthetic data. MIA measures privacy by checking resistance to inference attacks, ensuring an adversary cannot easily determine the membership status of data points in the training set. A lower score indicates better privacy.

Identifiability Score. This score measures the risk of re-identification by evaluating how easily individuals from the real

dataset can be identified in the synthetic dataset. It can assess privacy risk, ensuring sensitive information about individuals in the real dataset is not easily inferred from the synthetic data. A lower score indicates better privacy.

k-Map Score. The k-map score quantifies the number of real data points that share the same characteristics as a synthetic data point, providing a measure of anonymity. The minimum value k which satisfies the k-map rule: every combination of values for the quasi-identifiers appears at least k times in the reidentification (synthetic) dataset [19]. A higher score indicates better privacy.

3.3 Evaluation Synthetic data generation and evaluation process

The process for preparing, synthesising and evaluating the two synthetic data generation algorithms (i.e., GReaT and CTGAN) is described below in Figure 1. The workflow for synthetic data generation and evaluation starts with using raw datasets to train a synthetic data generation model. Next, the synthetic data generated is of the same number of rows as the original dataset. After that, synthetic data is evaluated based on resemblance (WD, JSD, and MMD), utility (AUROC scores with XGBoost, MLP, GMM), and privacy (MIA, identifiability score, k-map score). The evaluation is repeated twice to ensure robustness and reliability. This ensures that the synthetic data is evaluated in a comprehensive way.

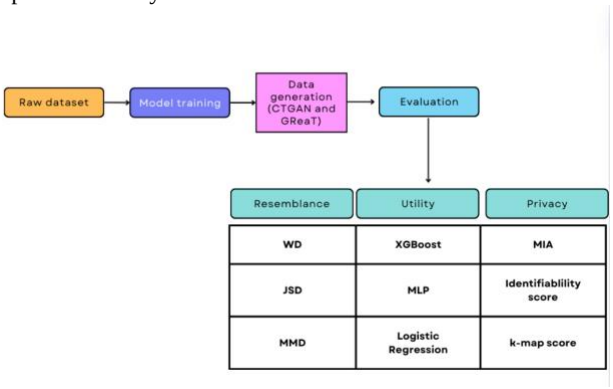

Figure 1: Workflow of the synthetic data generation and evaluation process

4 Evaluation results and discussion

4.1 Evaluation results

Resemblance. Table 2 compares the resemblance dimension of synthetic data generated by GReaT and CTGAN using WD, JSD, and MMD. CTGAN achieves lower values in WD (0.056475 vs. 0.059728) and MMD (0.010097 vs. 0.009997), indicating better performance in maintaining resemblance to real data. GReaT has a slightly lower JSD (0.008323 vs. 0.009633), suggesting it captures the categorical data distribution slightly better than CTGAN. However, in general, the performance of the two algorithms is very close to each other in the three metrics of the resemblance dimension, and the maximum difference is less than or equal to 0.003. Therefore, it can be summarised that in terms of resemblance performance, CTGAN and GReaT are highly comparable and demonstrate equivalent ability to mimic the real dataset.

Table 2: Resemblance dimension evaluation of GReaT and CTGAN

Synthetic data generation algorithms	WD	JSD	MMD
GReaT	0.059728	**0.008323**	**0.009997**
CTGAN	**0.056475**	0.009633	0.010097

Utility. Table 3 presents the utility dimension evaluation results of two synthetic data generation algorithms, GReaT and CTGAN, on different machine learning models (XGBoost, MLP, and Logistic Regression). The evaluation metric used is AUCROC (Area Under the Receiver Operating Characteristic Curve), which measures the classification performance of the models. Specifically, the higher the AUCROC value, the better the model performance. It can be seen that the baseline model has the highest AUCROC values, indicating that models trained and tested on real data perform the best. The synthetic data generated by GReaT achieves higher AUCROC values across all models compared to CTGAN, suggesting that the quality of synthetic data produced by GReaT is relatively higher and closer to real data. Overall, while synthetic data generation algorithms like GReaT and CTGAN can generate useful data to some extent, there is still a performance gap compared to models based on real data. Among them, GReaT performs relatively better, but there is still a need for improvement to narrow the gap with the baseline model.

Table 3: Utility dimension evaluation of GReaT and CTGAN

Synthetic data generation algorithms	XGBoost	MLP	GMM
Baseline model	0.772471	0.862595	0.871643
GReaT	**0.665974**	**0.697826**	**0.714905**
CTGAN	0.314789	0.410508	0.417898

Privacy. Table 4 shows the privacy evaluation results using MIA, Identifiability Score, and k-Map Score. CTGAN has a slightly lower MIA score (0.495062 vs. 0.504938) and Identifiability Score (0.31500 vs. 0.47500), indicating better privacy protection in these aspects. However, GReaT has a higher k-Map Score (2.00000 vs. 1.00000), suggesting it offers better protection against re-identification based on rare data characteristics. A higher k-Map Score indicates better privacy because it means every combination of quasi-identifiers appears at least k times, reducing the risk of re-identification.

Table 4: Privacy dimension evaluation of GReaT and CTGAN

Synthetic data generation	MIA	Identifiability Score	k-Map Score

algorithms			
GReaT	0.504938	0.47500	**2.00000**
CTGAN	**0.495062**	**0.31500**	1.00000

4.2 Discussion

From the experimental results of the case dataset, we summarise our findings as below:

Performance Across Three Dimensions: The LLM-based GReaT, designed specifically for tabular data generation, demonstrates performance comparable to the best performers in tabular data across all three dimensions, particularly excelling in machine learning tasks. This aligns with the findings of [2], who tested GReaT in other domains such as finance, real estate, and healthcare. However, this study extends beyond [2] by including the privacy dimension in the evaluation, whereas they only considered resemblance and utility. GReaT's performance in terms of privacy is also comparable to CTGAN, with each winning on different metrics of the three privacy metrics. Considering that CTGAN had previously demonstrated good privacy performance on educational tabular datasets compared to other generation algorithms [16], this further highlights the advancement of GReaT. Overall, LLM-based GReaT shows a balanced performance across resemblance, utility, and privacy dimensions.

No Universal Best Performer: Consistent with the findings of [20], no synthetic data generation method excels across all evaluation metrics. Our results indicate that GReaT, as an LLM, with its advanced semantic understanding and flexibility, holds significant promise in applications requiring detailed categorical data synthesis and certain privacy protections. Meanwhile, GAN-based CTGAN continues to excel in producing data with high predictive utility and robust resemblance to real data. Therefore, it remains essential to evaluate the performance of synthetic data generation methods on specific datasets to determine the most suitable approach for each use case.

5 Study limitations and future directions

One limitation of this study is the use of only one dataset for evaluation. This is because GReaT requires significantly more computation time compared to CTGAN, which limits the ability to compare more extensive datasets in experiments. Diversifying and expanding datasets in future studies would provide a more comprehensive representation of the algorithms' performance. Moreover, exploring a wider variety of datasets could help in understanding the strengths and weaknesses of each model in different contexts. As more LLMs are developed for synthetic tabular data generation in the future, it would be valuable to compare these new models against traditional high-performance algorithms.

6 Conclusion and Implications

Our study employs the advanced LLM-based GReaT and the top-performing CTGAN for tabular data generation, focusing on their performance in resemblance, utility, and privacy on educational tabular datasets. Our results demonstrate that the traditionally underestimated LLM-based synthetic data generation algorithms can match the best tabular data generation algorithms in the experimental education tabular dataset. This highlights the potential of LLMs in synthetic data generation, offering a viable alternative to traditional methods. In the possible future, as research progresses, some advantages of LLMs, such as requiring less preprocessing due to their ability to work directly with raw textual and categorical data, may lead to more advanced tabular data generation techniques. This could ultimately enhance the ability to generate realistic and useful synthetic data across various domains. Specifically, in the education sector, where there is a significant demand for tabular data but issues like student privacy and sample bias often result in limited and low-quality data, the potential of LLMs is particularly promising. It is important to note that while the current performance of LLM (GReaT) on tabular data mirrors that of one of the leading algorithms, CTGAN, further development of LLM-based synthetic tabular data generation algorithms is still required. Future research directions include experimenting with more diverse educational datasets and fine-tuning LLMs for tabular data. Nevertheless, we foresee significant potential benefits for the education discipline.

REFERENCES

[1] Ashrafi, N., Schmitt, V., Spang, R. P., Möller, S., & Voigt-Antons, J.-N. (2024). *Protect and Extend - Using GANs for Synthetic Data Generation of Time-Series Medical Records*. Arxiv.org. https://arxiv.org/html/2402.14042v1

[2] Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M., & Kasneci, G. (2023). LANGUAGE MODELS ARE REALISTIC TABULAR DATA GENERATORS. *The Eleventh International Conference on Learning Representations*. https://openreview.net/pdf?id=cEygmQNOeI

[3] Chang, Y., Wang, X., Wang, J., Yuan, W., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3). https://doi.org/10.1145/3641289

[4] Fang, X., Xu, W., Tan, F. A., Zhang, J., Hu, Z., Qi, Y., Nickleach, S., Socolinsky, D., Sengamedu, S., & Faloutsos, C. (2024). Large Language Models(LLMs) on Tabular Data: Prediction, Generation, and Understanding -- A Survey. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2402.17944

[5] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research*, *13*(25), 723–773. https://jmlr.csail.mit.edu/papers/v13/gretton12a.html

[6] Hernadez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2023). Synthetic Tabular Data Evaluation in the Health Domain Covering Resemblance, Utility, and Privacy Dimensions. *Methods of Information in Medicine*, *62*(Suppl 1), e19–e38. https://doi.org/10.1055/s-0042-1760247

[7] Hosbach, A. (2023). *LibGuides: Education Statistics, Data Sets, and Data Experts: Statistics and Data Sets*. Guides.lib.virginia.edu. https://guides.lib.virginia.edu/statisticsdatasets

[8] Hyland, S., Esteban, C., & Rätsch, G. (2017). REAL-VALUED (MEDICAL) TIME SERIES GENERA- TION WITH RECURRENT CONDITIONAL GANS. https://arxiv.org/pdf/1706.02633

[9] Islam, R. (2023). *Unveiling the Potential of CTGAN: Harnessing Generative AI for Synthetic Data*. KDnuggets. https://www.kdnuggets.com/2023/04/unveiling-potential-ctgan-harnessing-generative-ai-synthetic-data.html

[10] Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S., & Weller, A. (2022). *Synthetic Data -what, why and how?*

https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/Synthetic_Data_Survey-24.pdf

[11] Khosravi, H., Farhadpour, S., Grandhi, M., Raihan, A. S., Das, S., & Ahmed, I. (2023, November 15). *Strategic Data Augmentation with CTGAN for Smart Manufacturing: Enhancing Machine Learning Predictions of Paper Breaks in Pulp-and-Paper Production*. ArXiv.org. https://doi.org/10.48550/arXiv.2311.09333

[12] Kim, J., Kim, T., & Choo, J. (2024). *Group-wise Prompting for Synthetic Tabular Data Generation using Large Language Models*. https://arxiv.org/pdf/2404.12404

[13] Kimmons, R. (2018). Students Performance in Exams. *Www.kaggle.com*. https://www.kaggle.com/datasets/spscientist/students-performance-in-exams

[14] Kolouri, S., Park, S., Thorpe, M., Slepčev, D., & Rohde, G. K. (2017). Optimal Mass Transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, *34*(4), 43–59. https://doi.org/10.1109/MSP.2017.2695801

[15] Liu, Q., & Khalil, M. (2023). Understanding privacy and data protection issues in learning analytics using a systematic review. *British Journal of Educational Technology*, *54*(6). https://doi.org/10.1111/bjet.13388

[16] Liu, Q., Khalil, M., Jovanovic, J., & Shakya, R. (2024, March 18). Scaling While Privacy Preserving: A Comprehensive Synthetic Tabular Data Generation and Evaluation in Learning Analytics. *Proceedings of the 14th Learning Analytics and Knowledge Conference*. https://doi.org/10.1145/3636555.3636921

[17] Menéndez, M. L., Pardo, J. A., Pardo, L., & Pardo, M. C. (1997). The Jensen-Shannon divergence. *Journal of the Franklin Institute*, *334*(2), 307–318. https://doi.org/10.1016/s0016-0032(96)00063-4

[18] Nikitin, A. (2023). *Evaluation of Synthetic Time Series*. Medium. https://towardsdatascience.com/evaluation-of-synthetic-time-series-1b4fc4e2be39

[19] Qian, Z., Cebere, B.-C., & Mihaela, van der S. (2023). Synthcity: facilitating innovative use cases of synthetic data in different data modalities. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2301.07573

[20] Rosenblatt, L., Liu, X., Pouyanfar, S., de Leon, E., Desai, A., & Allen, J. (2020, November 10). Differentially Private Synthetic Data: Applied Evaluations and Enhancements. ArXiv.org. https://doi.org/10.48550/arXiv.2011.05537

[21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). *Attention Is All You Need*. ArXiv.org. https://arxiv.org/abs/1706.03762

[22] Xu, L., Skoularidou, M., Cuesta, A., & Veeramachaneni, K. (2019). *Modeling Tabular Data using Conditional GAN*. https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf

[23] Ye, H., Batbaatar, E., Choi, D.-W., Choi, K. S., Ko, M., & Ryu, K. S. (2023). Synthetic Tabular Data Based on Generative Adversarial Networks in Health Care: Generation and Validation Using the Divide-and-Conquer Strategy. *JMIR Medical Informatics*, *11*, e47859–e47859. https://doi.org/10.2196/47859

[24] Zhao, Z., Kunar, A., Birke, R., & Chen, L. (2021). CTAB-GAN: Effective Table Data Synthesizing. *Proceedings of Machine Learning Research*, *157*, 2021–2021. https://proceedings.mlr.press/v157/zhao21a/zhao21a.pdf