SLATE
Centre for the Science of Learning & Technology

UNIVERSITAS BERGENSIS

Welcome to the world of synthetic data

Exploring the Generation of synthetic educational tabular data using LLMs

26th August 2024

KDD2024
BARCELONA, SPAIN

Authors

Qinyi Liu

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

Mohammad Khalil

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

KDD2024
BARCELONA, SPAIN

We are motivated by

1. Tabular data is critical across various fields, including education, forming the backbone of numerous classification and regression machine learning applications for education data science.

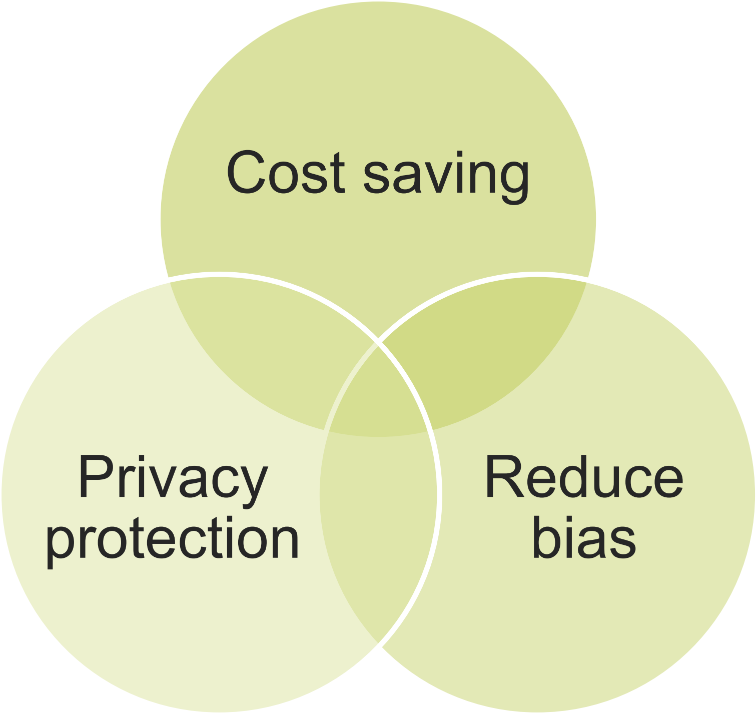
2. In practical applications, obtaining sufficient and high-quality tabular data often faces multiple challenges, including high costs, time-consuming processes, and strict data privacy protection requirements.

3. **Synthetic data generation has emerged as a promising solution!**

KDD2024 BARCELONA, SPAIN

What is synthetic data?

Synthetic data can



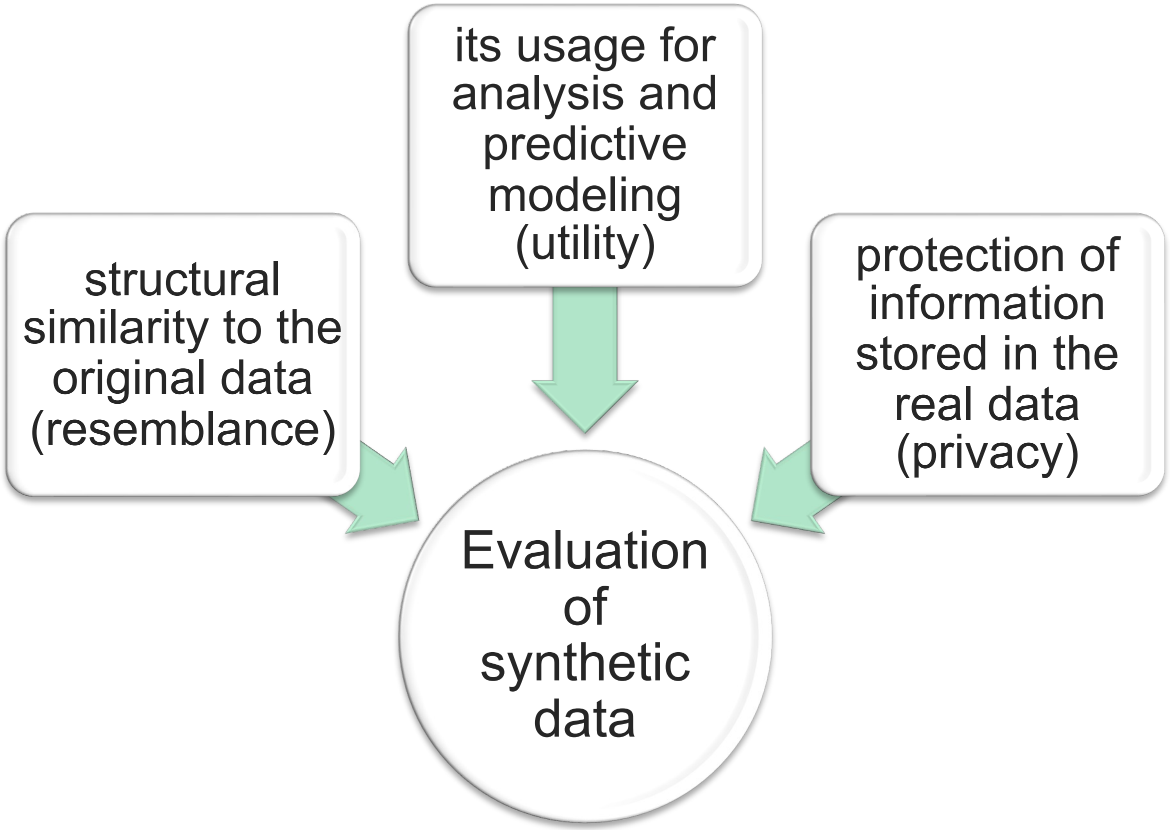
“Synthetic data is defined as the artificially annotated information generated by computer algorithms or simulations” [1].

KDD2024
BARCELONA, SPAIN

More about synthetic data: generation and evaluation

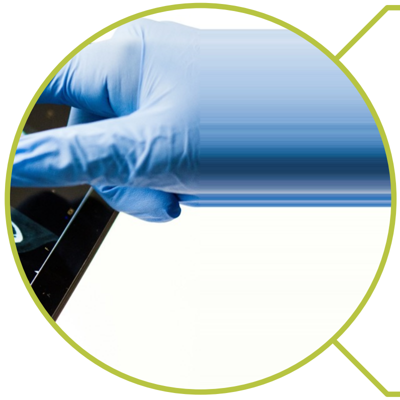
1. The most common synthetic tabular data generation methods can be divided into two types, statistical distribution methods and deep learning-based methods.

2. Large Language Models (LLMs) have emerged as one of the promising approaches in deep learning. Nevertheless, LLMs have not yet demonstrated excellence in producing high-quality synthetic tabular data.



KDD2024
BARCELONA, SPAIN

More about synthetic data- application



For application, synthetic data has been widely used in health, robotics training, transportation sign detection field

Only a few application of synthetic data in **education field**, they tend to consider only some of the evaluation dimensions

KDD2024
BARCELONA, SPAIN

What we did?

Generation of Realistic Tabular data
with pretrained Transformer-based language models

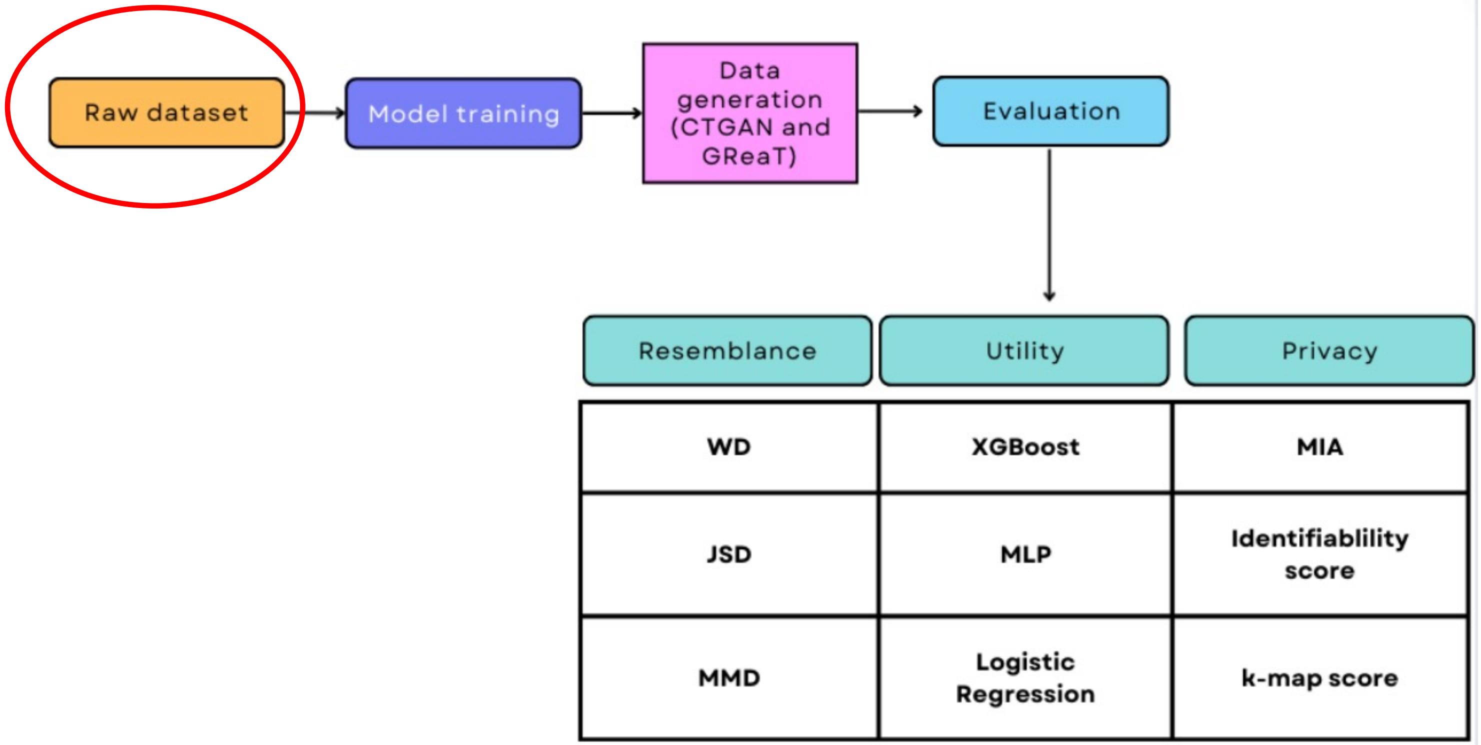
We evaluate GReaT, an advanced LLM for tabular data generation [2], and compare it with CTGAN on educational datasets.

Our contribution

- Our research bridges the gap in applying LLMs to educational tabular data generation, showcasing advanced techniques and experiments that enhance educational data analytics.
- Our findings demonstrate that GReaT and CTGAN perform comparably across multiple metrics, suggesting promising future applications of LLMs in educational data..

KDD2024
BARCELONA, SPAIN

Methodology



KDD2024
BARCELONA, SPAIN

Dataset and environment

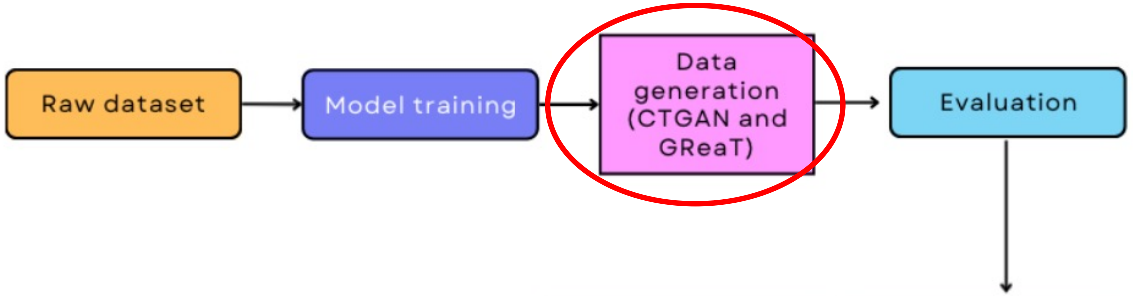
Table 1: Details of the used dataset

Dataset character	**Dataset detail**
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

Environment: python library called Synthcity [3], in Google Colab with NVIDIA A100

KDD2024 BARCELONA, SPAIN

Methodology



Resemblance	Utility	Privacy
WD	**XGBoost**	**MIA**
JSD	**MLP**	**Identifiablility score**
MMD	**Logistic Regression**	**k-map score**

Synthetic data generators

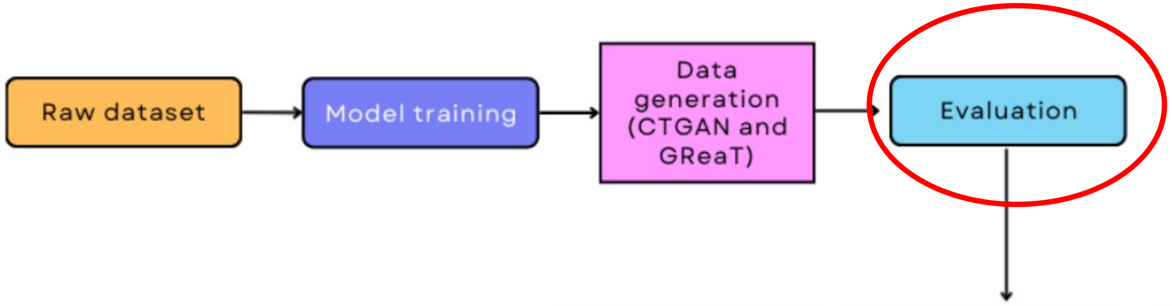
CTGAN (Conditional Tabular GAN) is a generative adversarial network designed specifically for generating synthetic tabular data [4]. It handles complex data distributions and categorical variables, making it popular for tasks requiring realistic synthetic datasets in structured data formats.

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

GReaT [2] is an advanced LLM tailored for tabular data generation. It leverages the capabilities of LLMs to produce high-quality synthetic data, with a focus on maintaining resemblance and utility, making it a promising tool for applications in domains like health. <u>But it has not been test in education domain dataset.</u>

KDD2024
BARCELONA, SPAIN

Methodology



Resemblance	Utility	Privacy
WD	XGBoost	MIA
JSD	MLP	Identifiablility score
MMD	Logistic Regression	k-map score

Evaluation metrics

Resemblance

- (Average Jensen-Shannon Distance) JSD: measures the similarity between two probability distributions, especially for categorical data, lower is better
- Maximum Mean Discrepancy (MMD): measure the distributions in kernel Hilbert space (RKHS), for categorical and continuous data, lower is better
- Wasserstein Distance (WD): also known as the Earth Mover's Distance, , especially for continuous data, lower is better

Utility

- Validating synthetic data utility in downstream machine learning task
- Train-Synthetic-Test-Real
- Logistic Regression, Multi-Layer Perceptron (MLP), and Extreme Gradient Boosting (XGBoost)

Privacy

- Membership Inference Attack (MIA) with Prior Knowledge: lower is better
- Identifiability Score: lower is better
- k-Map Score: higher is better

Findings – resemblance and utility

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

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

KDD2024 BARCELONA, SPAIN

Findings - privacy

Table 4: Privacy dimension evaluation of GReaT and CTGAN

Synthetic data generation algorithms	**MIA**	**Identifiability Score**	**k-Map Score**
GReaT	0.504938	0.47500	**2.00000**
CTGAN	**0.495062**	**0.31500**	1.00000

Discussion and limitations

1. 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.

2. No Universal Best Performer

Limitations:

Only single dataset used for evaluation- as GReaT requires significantly more computation time and resources compared to CTGAN

Conclusion and implications

- Results show that LLMs, like GReaT, can match top algorithms like CTGAN, highlighting their potential in synthetic data generation.

- LLMs offer advantages such as reduced preprocessing, especially with raw textual and categorical data, which could lead to more advanced tabular data generation techniques in the future.

- The potential of LLMs for synthetic data generation is particularly promising in education, where privacy concerns and data limitations are prevalent.

- Future research should explore diverse educational datasets and fine-tune LLMs for tabular data to further enhance their performance.

KDD2024 BARCELONA, SPAIN

Future direction

Build a user-friendly synthetic data generation platform specifically for education domain (www.lasd.ai)

Thank you & references

Link to paper

If you want to contact me: qinyi.liu@uib.no

[1] 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

[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] 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

[4] 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

KDD2024 BARCELONA, SPAIN

THANK YOU

KDD2024
BARCELONA, SPAIN