NEURAL INFORMATION PROCESSING SYSTEMS

HARMONIC: Harnessing LLMs for Tabular Data Synthesis and Privacy Protection

Yuxin Wang, Duanyu Feng, Yongfu Dai, Zhengyu Chen, Jimin Huang, Sophia Ananiadou, Qianqian Xie, Hao Wang

SICHUAN UNIVERSITY

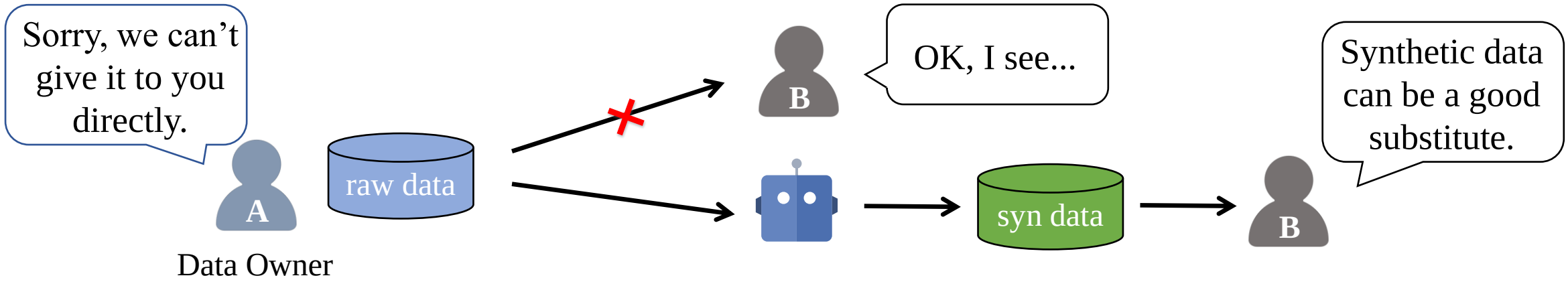
WUHAN UNIVERSITY

MANCHESTER 1824
The University of Manchester

THE FIN AI
FINANCIAL ARTIFICIAL INTELLIGENCE

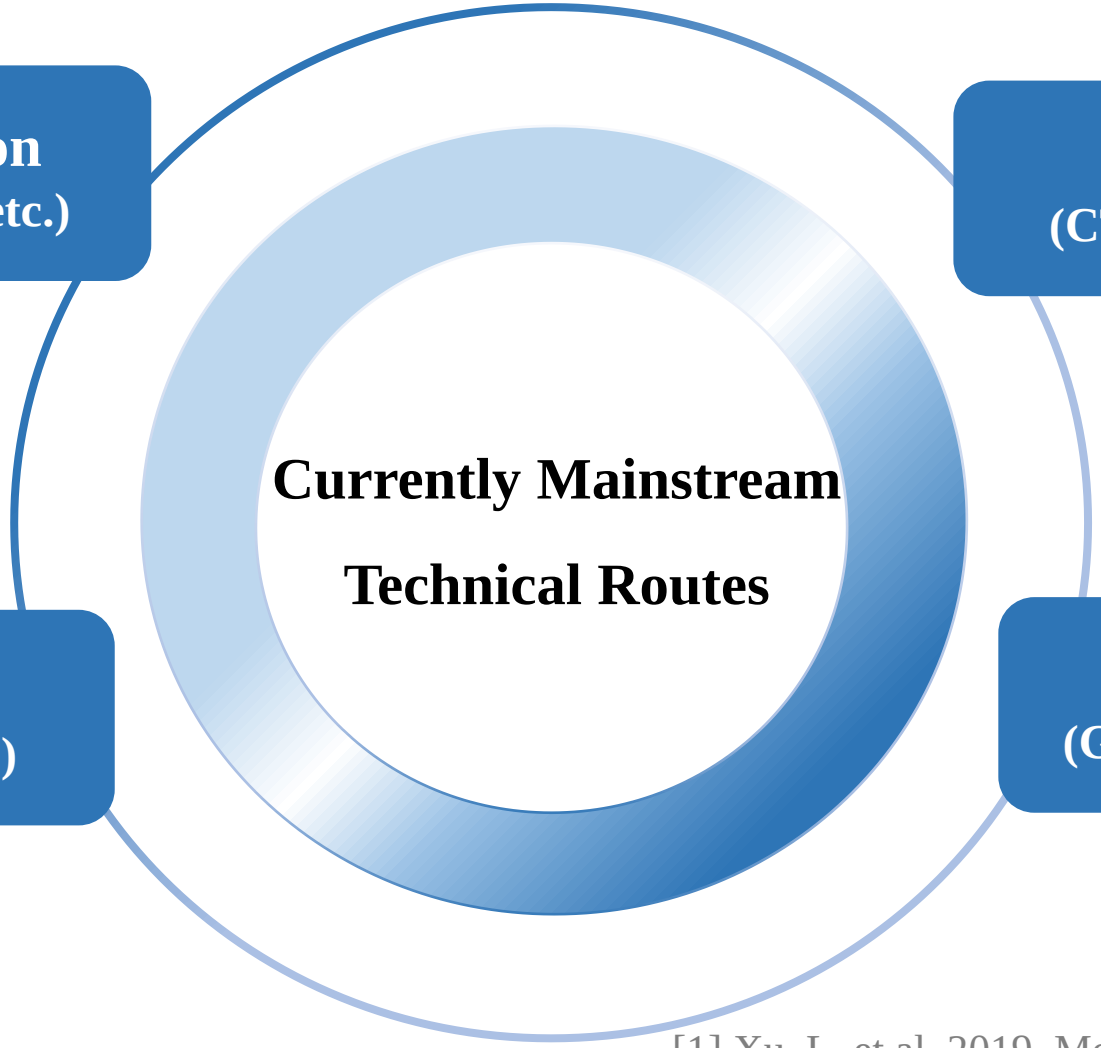
Data Synthesis (What is/for)

- Synthetic data is defined as data obtained from a generative process that **learns the properties of the original data**.
- It intends to synthesize new samples that are related to but can **not be mapped back to the original data**.
- The generation of synthetic data can be used for **anonymization**, **regularization**, **oversampling**, **semi-supervised learning**, and several other tasks.



[1] Assefa, S. A. et al. 2022. Generating synthetic data in finance: opportunities, challenges and pitfalls.
[2] Fonseca, J. et al. 2023. Tabular and latent space synthetic data generation: a literature review.

Data Synthesis (How to do)



Simple Augmentation
(SMOTE、FW-SMOTE, etc.)

easy, but overlook semantic information

GAN-based
(CTGAN、GTAB-GAN, etc.)

good privacy, but data quality is average

Diffusion-based
(TabDDPM、TABSYN, etc.)

easy, but overlook semantic information

LLM-based
(GReaT、REaLTabFormer, etc.)

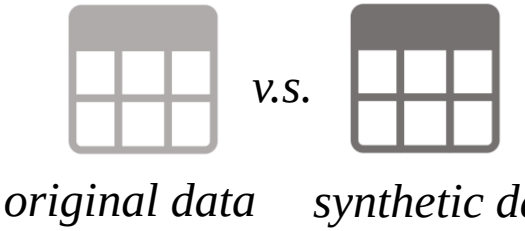
effective, but risky for data leakage

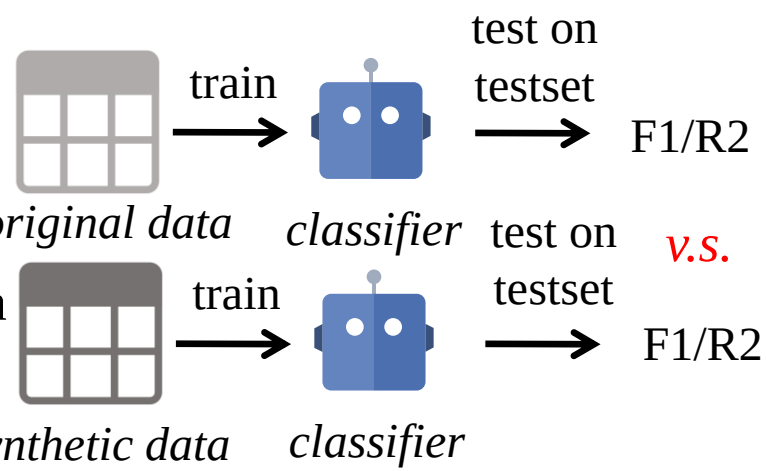
[1] Xu, L. et al. 2019. Modeling tabular data using conditional gan.
[2] Chawla, N. V. et al. 2002. SMOTE: synthetic minority over-sampling technique.
[3] Zhang, H. et al. 2023. Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space.
[4] Solatorio, A. V. et al. 2023. REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers.

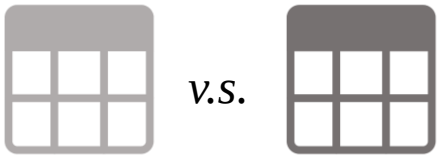
Data Synthesis (How to evaluate?)

Metrics

- **Effectiveness**
  - Statistical Characteristics
    - Compare distribution similarity
    - data_mismatch, Wasserstein_dist, CorrelationSimilarity
  - Peformanc on Downstream Task
    - Train on original data v.s. Train on synthetic data
    - Machine Learning Efficiency (MLE)
- **Privacy**
  - Measure the distance between synthetic data and original data
  - Distance to Closest Record (DCR), NewRowSynthesis







[1] DataCebo, Inc. Synthetic Data Metrics, 10 2023. Version 0.12.0.
[2] Xu, L. et al. 2019. Modeling tabular data using conditional gan.
[3] Zhao, Z. et al. 2021. Ctab-gan: Effective table data synthesizing.

HARMONIC: Harnessing LLMs for Tabular Data Synthesis and Privacy Protection

Generation

Purpose:

Improve privacy while preserving the efficacy of LLMs

Evaluation

Purpose:

Evaluate the effectiveness and privacy risks of synthetic data in the age of LLMs.

HARMONIC: Harnessing LLMs for Tabular Data Synthesis and Privacy Protection

Generation

Propose a synthetic tabular data generation approach, which utilizes:

I. KNN to maintain the efficacy

II. fine-tuning LLMs for privacy preservation

Evaluation

Introduce two new metrics to evaluate the **effectiveness** and **privacy** of synthetic data for LLM-based synthesis methods:

I. LLM Effectiveness (LLE)

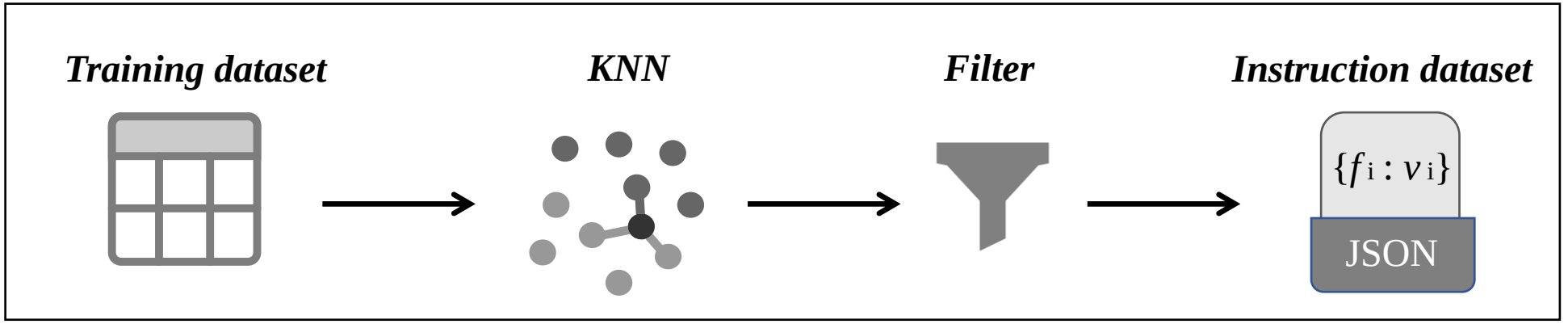
II. Data Leakage Test (DLT)

HARMONIC: Harnessing LLMs for Tabular Data Synthesis and Privacy Protection

Generation

◉ **I. Instruction dataset construction**

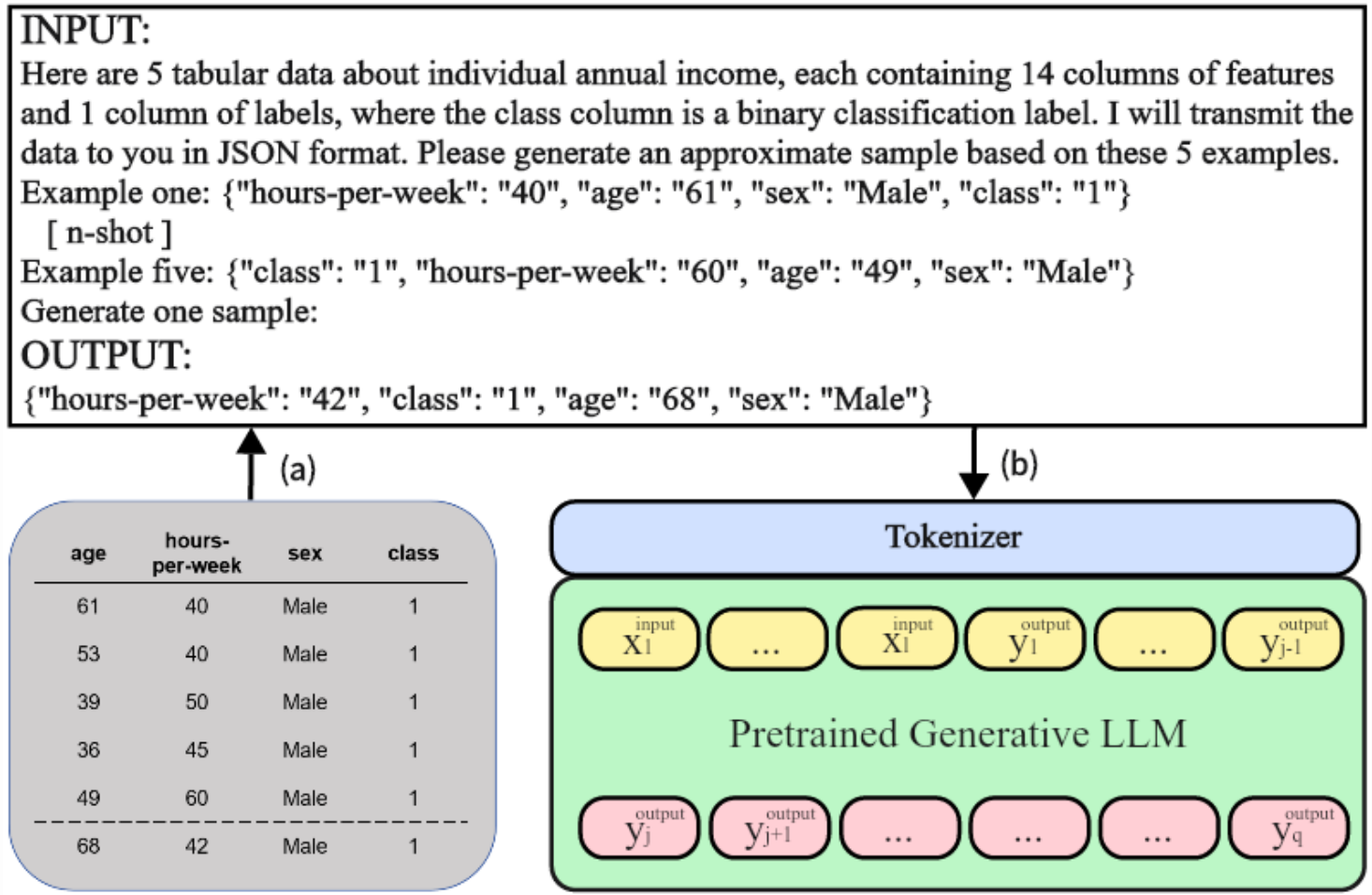
- Let the LLMs see the relationship between multiple similar rows and construct the structural tabular synthetic data format.
- To achieve this, we use the **KNN** algorithm to identify neighboring data for each instance.
- Every row of tabular data set getted by KNN is converted into **JSON** dictionary format.



HARMONIC: Harnessing LLMs for Tabular Data Synthesis and Privacy Protection

Generation

- **II. Instruction tuning based tabular data synthesizer**
  - Fine-tune the LLM for the synthetic data generation task using the instruction dataset.
  - The **objective** of our fine-tuning strategy is to maximize the probability of generating the correct output sequence given the prompt describing the task and 5 input real data points.

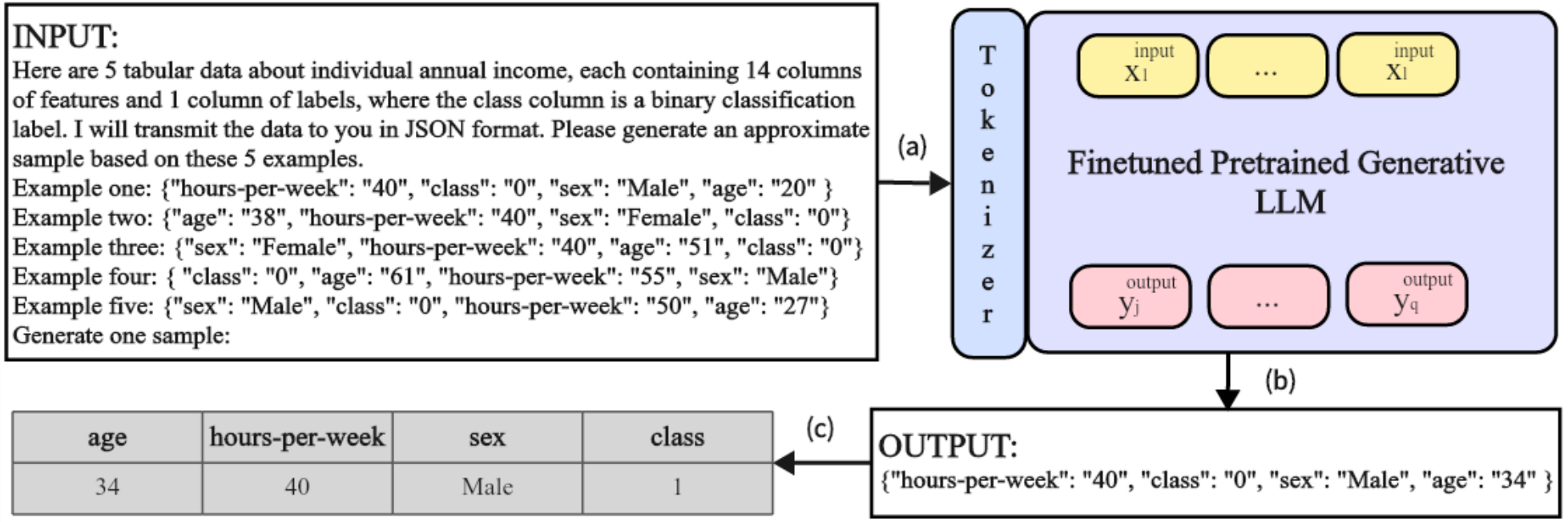


HARMONIC: Harnessing LLMs for Tabular Data Synthesis and Privacy Protection

Generation

◉ **III. Sampling for synthetic data generation**

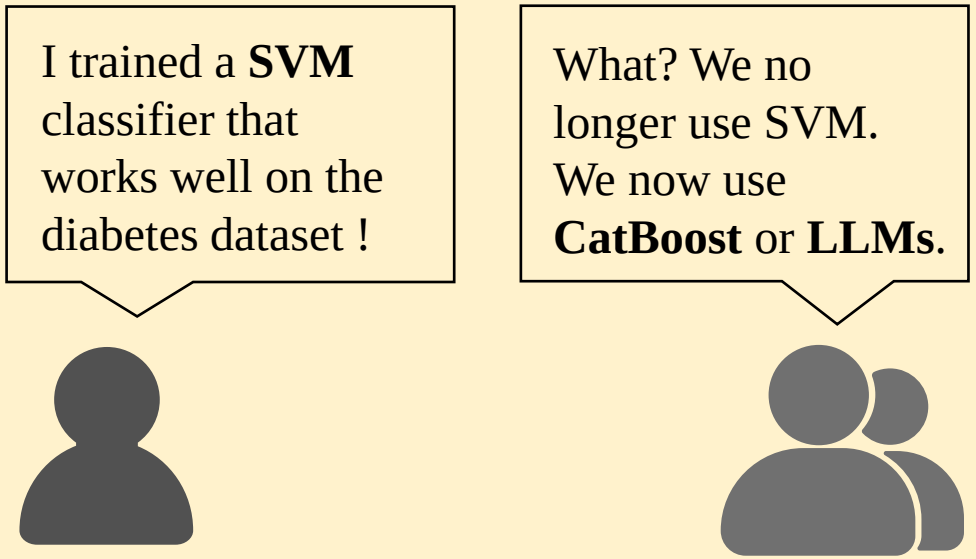
- After the fine-tune, we denote the fine-tuned LLM as the generator for tabular data synthesis.
- The prompt dataset used for data generation is re-sampled randomly and is disjoint from the training set.
- Our method aims to teach the LLMs to extract patterns from the original data, unlike pretraining where the model primarily memorizes the data.

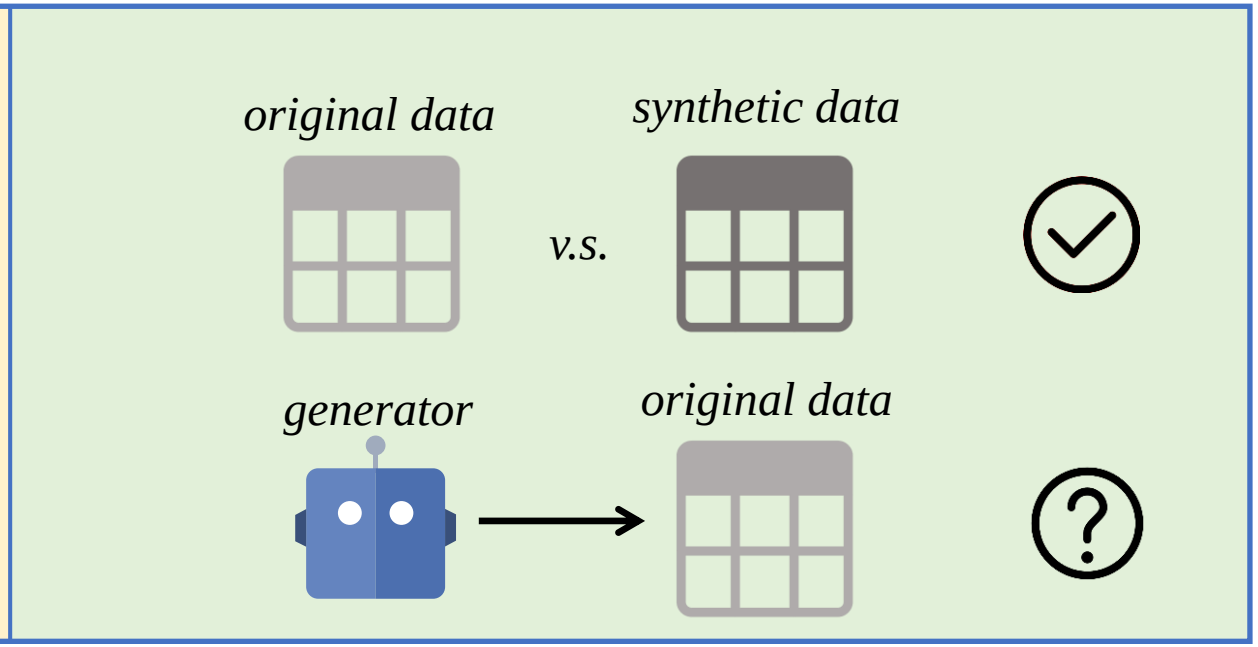


HARMONIC: Harnessing LLMs for Tabular Data Synthesis and Privacy Protection

Evaluation

- Motivation
  - With the development of LLMs, evaluating the effectiveness of synthetic data using weak classifiers(MLE) is losing its practical value and credibility.
  - The commonly used data leakage metrics focus on measuring the "distance" between synthetic data and real data, without taking into account the extent to which the generator itself leaks data.





HARMONIC: Harnessing LLMs for Tabular Data Synthesis and Privacy Protection

Evaluation

LLE

- We propose using synthetic data to fine-tune a pretrained LLM and then evaluate the fine-tuned LLM on the real test set. We refer to this as LLM Effectiveness (LLE).
- LLMs trained on synthetic data should ideally achieve comparable or even surpass the performance of LLMs trained on real data, as measured by LLE on a real test set.

DLT

- We introduce a new metric, Data Leakage Test (DLT), to quantify privacy by comparing the difference in perplexity (PPL) between a generator on real data and synthetic data as:

$$DLT = PPL(D_{train}) - PPL(D_{syn})$$

- A higher DLT value indicates that the generator is more inclined to generate synthetic data rather than real data, thus reducing the likelihood of leaking real data.

Experimental setup

- **Task & Data**: four classification datasets (German, Adult Income, Diabetes, Buddy)

- **Baselines:** Simple Augmentation (SMOTE), VAE-based (TVAE), GAN-based (CTABGAN), Diffusion-based (TabDDPM, TABSYN), LLM-based (REaLTabFormer, GReaT)

- **Evaluation Metric**: statistical characteristics (data_mismatch, Wasserstein_dist, CorrelationSimilarity), effectiveness (MLE, LLE), privacy (NewRowSynthesis, DCR, DLT)

- **Models**: we opt for LLaMA-2-7b-chat as the base model

- **Ablations**:
  - *w/o KNN*: Using randomly sampling instead of KNN in Instruction Dataset Construction
  - *w/o filter*: Don't use the filter operation in Instruction Dataset Construction
  - *w/o permutation*: Don't permute the features in Instruction Dataset Construction

[1] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994.
[2] https://github.com/vanderschaarlab/synthcity
[3] Hugo Touvron, et al. Llama 2: Open foundation and fine-tuned chat models.

Experiments – Main Results

Dataset	Metric	Original	HARMONIC	SMOTE	TVAE	CTAB	TabDDPM	TABSYN	RTF	GReaT
GM	DM	–	**$0.00_{\pm 0.00}$**	$0.14_{\pm 0.00}$	$0.14_{\pm 0.00}$	$0.14_{\pm 0.00}$	$0.14_{\pm 0.00}$	$0.27_{\pm 0.00}$	**$0.00_{\pm 0.00}$**	$0.14_{\pm 0.00}$
	WD	–	$0.87_{\pm 0.07}$	$0.85_{\pm 0.03}$	$0.70_{\pm 0.04}$	$0.77_{\pm 0.02}$	$0.73_{\pm 0.02}$	$0.94_{\pm 0.09}$	**$0.67_{\pm 0.01}$**	$0.93_{\pm 0.22}$
	CS	–	0.96	0.97	**0.99**	0.98	0.90	0.98	0.98	0.98
	MLE	$0.50_{\pm 0.00}$	$0.55_{\pm 0.03}$	$0.64_{\pm 0.02}$	$0.61_{\pm 0.02}$	$0.57_{\pm 0.02}$	$0.64_{\pm 0.01}$	$0.63_{\pm 0.02}$	**$0.65_{\pm 0.01}$**	$0.44_{\pm 0.03}$
	LLE	$0.71_{\pm 0.00}$	$0.64_{\pm 0.03}$	$0.67_{\pm 0.04}$	$0.69_{\pm 0.03}$	$0.71_{\pm 0.02}$	$0.67_{\pm 0.05}$	**$0.72_{\pm 0.02}$**	$0.69_{\pm 0.03}$	$0.55_{\pm 0.11}$
AD	DM	–	$0.21_{\pm 0.15}$	**$0.00_{\pm 0.00}$**	**$0.00_{\pm 0.00}$**	**$0.00_{\pm 0.00}$**	**$0.00_{\pm 0.00}$**	**$0.00_{\pm 0.00}$**	**$0.00_{\pm 0.00}$**	$0.06_{\pm 0.00}$
	WD	–	$0.48_{\pm 0.15}$	$0.49_{\pm 0.01}$	$0.31_{\pm 0.04}$	$0.07_{\pm 0.01}$	$0.06_{\pm 0.01}$	$0.07_{\pm 0.01}$	**$0.03_{\pm 0.00}$**	$3.83_{\pm 0.09}$
	CS	–	0.90	**0.99**	0.98	0.97	**0.99**	0.97	**0.99**	0.94
	MLE	$0.61_{\pm 0.00}$	$0.67_{\pm 0.02}$	$0.75_{\pm 0.00}$	$0.74_{\pm 0.00}$	$0.73_{\pm 0.01}$	$0.74_{\pm 0.00}$	$0.73_{\pm 0.01}$	**$0.76_{\pm 0.00}$**	$0.73_{\pm 0.01}$
	LLE	$0.81_{\pm 0.00}$	$0.80_{\pm 0.02}$	$0.84_{\pm 0.01}$	$0.83_{\pm 0.01}$	$0.83_{\pm 0.00}$	$0.83_{\pm 0.00}$	$0.81_{\pm 0.02}$	**$0.85_{\pm 0.00}$**	$0.82_{\pm 0.02}$
DI	DM	–	**$0.03_{\pm 0.05}$**	$0.07_{\pm 0.05}$	$0.10_{\pm 0.00}$	$0.10_{\pm 0.00}$	$0.07_{\pm 0.05}$	$0.10_{\pm 0.00}$	$0.10_{\pm 0.00}$	$0.07_{\pm 0.05}$
	WD	–	$0.14_{\pm 0.01}$	**$0.07_{\pm 0.00}$**	**$0.07_{\pm 0.00}$**	$0.26_{\pm 0.01}$	$0.08_{\pm 0.00}$	$0.09_{\pm 0.01}$	$0.09_{\pm 0.02}$	$0.13_{\pm 0.00}$
	CS	–	0.95	0.96	0.98	0.97	**0.99**	0.98	**0.99**	0.88
	MLE	$0.56_{\pm 0.00}$	$0.46_{\pm 0.02}$	**$0.72_{\pm 0.03}$**	$0.71_{\pm 0.02}$	$0.67_{\pm 0.02}$	$0.71_{\pm 0.02}$	$0.68_{\pm 0.03}$	$0.66_{\pm 0.03}$	$0.45_{\pm 0.03}$
	LLE	$0.70_{\pm 0.00}$	$0.75_{\pm 0.00}$	$0.69_{\pm 0.04}$	$0.72_{\pm 0.04}$	$0.62_{\pm 0.09}$	$0.72_{\pm 0.03}$	**$0.77_{\pm 0.01}$**	$0.70_{\pm 0.04}$	$0.71_{\pm 0.03}$
BU	DM	–	**$0.00_{\pm 0.00}$**	**$0.00_{\pm 0.00}$**	**$0.00_{\pm 0.00}$**	**$0.00_{\pm 0.00}$**	**$0.00_{\pm 0.00}$**	**$0.00_{\pm 0.00}$**	**$0.00_{\pm 0.00}$**	$0.03_{\pm 0.04}$
	WD	–	$0.48_{\pm 0.16}$	$0.23_{\pm 0.02}$	$0.10_{\pm 0.01}$	$0.05_{\pm 0.00}$	$0.06_{\pm 0.02}$	**$0.04_{\pm 0.00}$**	**$0.04_{\pm 0.00}$**	$2292.47_{\pm 1014.22}$
	CS	–	0.93	0.98	0.97	0.99	**1.00**	**1.00**	0.98	**1.00**
	MLE	$0.38_{\pm 0.00}$	**$0.27_{\pm 0.03}$**	$0.25_{\pm 0.02}$	**$0.27_{\pm 0.03}$**	$0.26_{\pm 0.01}$	**$0.27_{\pm 0.01}$**	$0.26_{\pm 0.01}$	$0.26_{\pm 0.00}$	$0.24_{\pm 0.03}$
	LLE	$0.88_{\pm 0.00}$	$0.82_{\pm 0.03}$	$0.85_{\pm 0.04}$	**$0.86_{\pm 0.01}$**	$0.82_{\pm 0.02}$	$0.85_{\pm 0.01}$	**$0.86_{\pm 0.01}$**	$0.70_{\pm 0.14}$	$0.81_{\pm 0.03}$

- Our method exhibits performance comparable to existing SOTA approaches, especially from a statistical perspective, our method approaches optimality.
- In the era of LLMs, detecting synthetic data on LLMs is essential but it doesn't necessarily have to align with the performance of machine learning models.

Experiments – Main Results

Dataset	Metric	HARMONIC	SMOTE	TVAE	CTAB	TabDDPM	TABSYN	RTF	GReaT
GM	NRS	**1.00**	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	DCR	**8.08**	2.77	4.09	5.36	2.21	3.98	4.60	5.84
	DLT	**-0.16**	—	—	—	—	—	-22.04	-2.14
AD	NRS	**1.00**	0.95	1.00	1.00	1.00	1.00	1.00	1.00
	DCR	**2.47**	0.16	0.49	0.82	0.50	0.86	0.57	1.51
	DLT	-0.98	—	—	—	—	—	-163.71	**-0.67**
DI	NRS	**1.00**	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	DCR	0.44	0.28	0.33	0.72	0.21	**1.37**	0.36	1.36
	DLT	**-0.37**	—	—	—	—	—	-42.46	-0.44
BU	NRS	**1.00**	0.93	1.00	1.00	0.99	1.00	1.00	1.00
	DCR	2.52	0.15	0.66	0.70	0.18	1.38	0.38	**8.30**
	DLT	**-0.34**	—	—	—	—	—	-41.13	-2.22

- Our method prioritizes privacy in the synthetic data generation, HARMONIC surpasses or comes in a close second for all datasets across all metrics.
- DLT indicator suggests that LLMs pose a significant risk of data leakage.

Conclusion

Contributions

- We recognize that it is crucial to not only focus on the strong data generation ability of LLM in this era, but also pay attention to the potential privacy risks it may bring.
- We develop a framework, HARMONIC, for tabular data synthesis based on LLM. Our method ensures privacy preservation while maintaining the effectiveness of synthetic data.
- Under the HARMONIC framework, a set of metrics is proposed for the effectiveness in downstream LLMs tasks and privacy risk evaluation of synthetic tabular data.

Future Work

- We will explore tasks beyond categorical datasets.
- We will also investigate data formats beyond tabular data.
- We will attempt to implement our methods in real-world applications.

Refer to our paper for more experiments and discussions

NEURAL INFORMATION PROCESSING SYSTEMS

HARMONIC: Harnessing LLMs for Tabular Data Synthesis and Privacy Protection

Yuxin Wang, Duanyu Feng, Yongfu Dai, Zhengyu Chen, Jimin Huang, Sophia Ananiadou, Qianqian Xie, Hao Wang

SICHUAN UNIVERSITY

WUHAN UNIVERSITY

MANCHESTER 1824
The University of Manchester

THE FIN AI
FINANCIAL ARTIFICIAL INTELLIGENCE