Generative AI Uses for Synthetic Data in Education

Privacy and performance in educational research

Mohammad Khalil

Associate Professor of AI and Education

Centre for the Science of Learning & Technology (SLATE), AI LEARN

University of Bergen, Norway

OECD | March 2026 | Online

Dystopia:

Privacy & Data Protection issues

Universities are tracking their students. Is it clever or creepy?

Learning analytics are becoming increasingly popular for improving learning and cutting drop-out rates – but critics question the impact on privacy

▲ 'Some students are concerned about us continuously monitoring in a Big Brother fashion.' Photograph: Alamy

The Data Challenge in Educational Research

Privacy Barriers

GDPR and data protection regulations restrict collection, storage, and sharing of student data across institutions and borders.

Data Scarcity

Predictive models use small samples. Insufficient data limits scalability and generalisability of research.

Quality & Bias

Incomplete datasets, collection biases, and class imbalances undermine the validity of educational models and interventions.

Synthetic data generation offers a promising path to address these interconnected challenges.

Privacy

When 'Anonymous' is still you

Different forms of anonymisation

useful
but not private

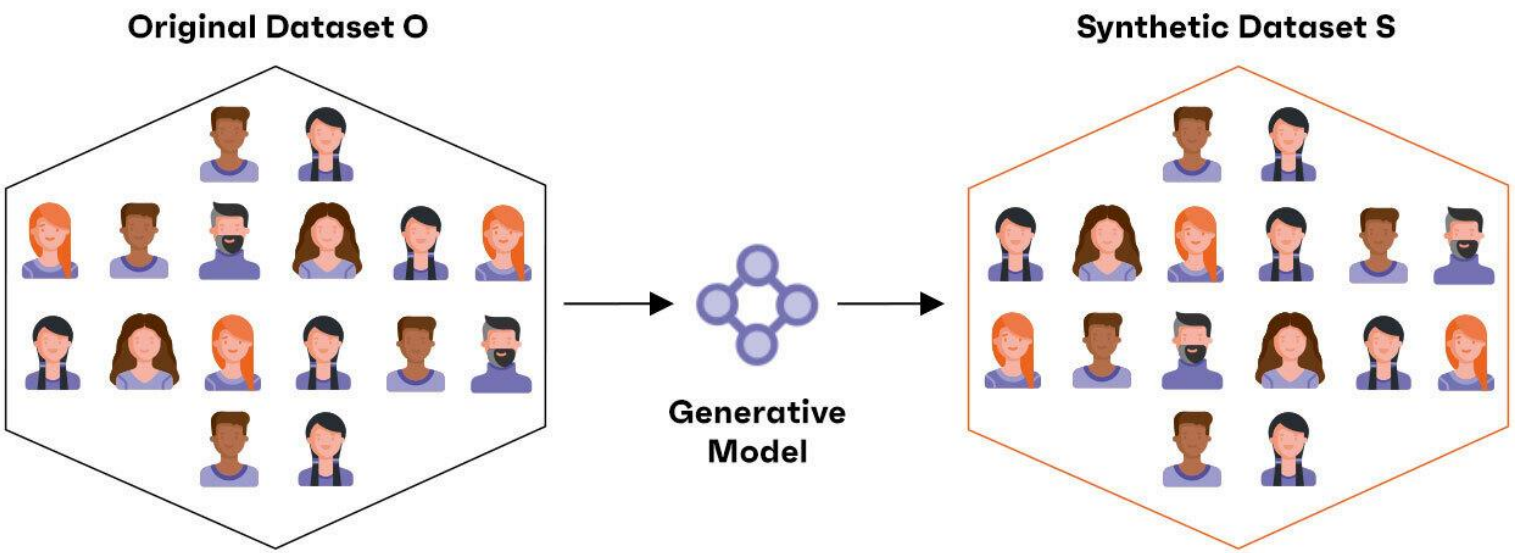
What is Synthetic Data?

Artificially generated data that mimics real-world data while preserving statistical properties.

The Spectrum of Synthetic Data

01

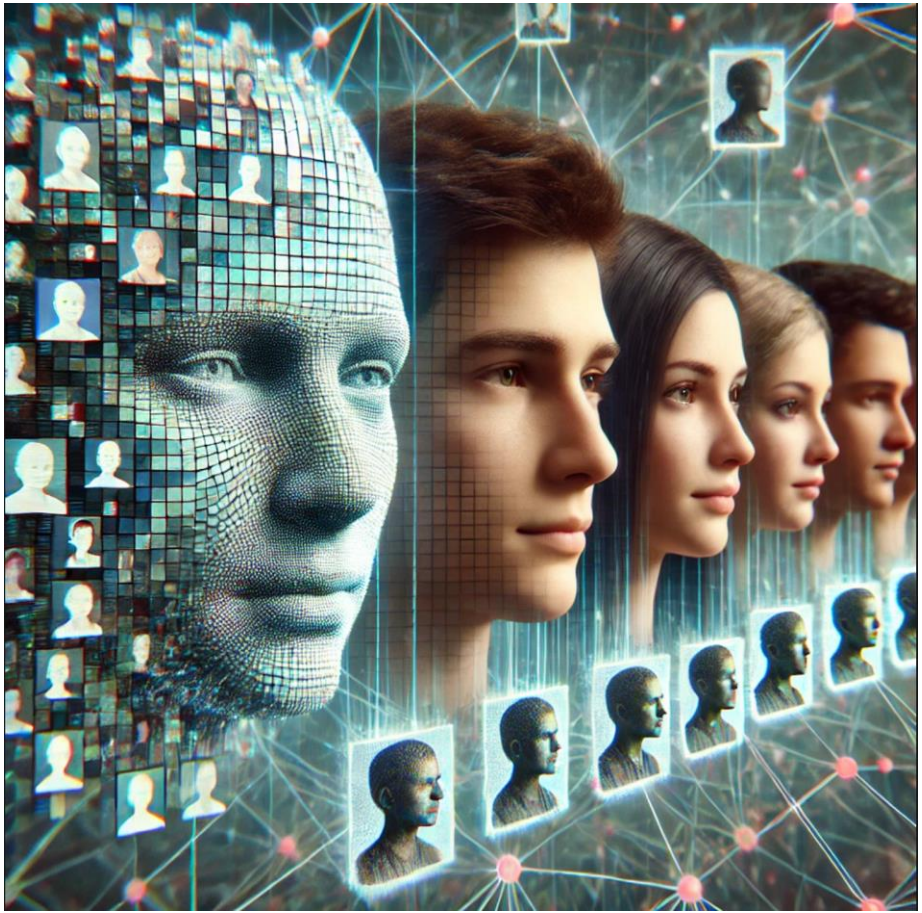
Synthetic Data for Privacy Preservation


Original Dataset O
Generative Model
Synthetic Dataset S

02

Synthetic Data for Performance & Scalability

Artificial Students that resemble actual students?



Tabular and time-series perspective*

Khalil, M., Vadiee, F., Shakya, R., & Liu, Q. (2025). Creating artificial students that never existed: Leveraging large language models and CTGANs for synthetic data generation. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 439-450).

Comparison of Real and Synthetic Data for Dataset B2

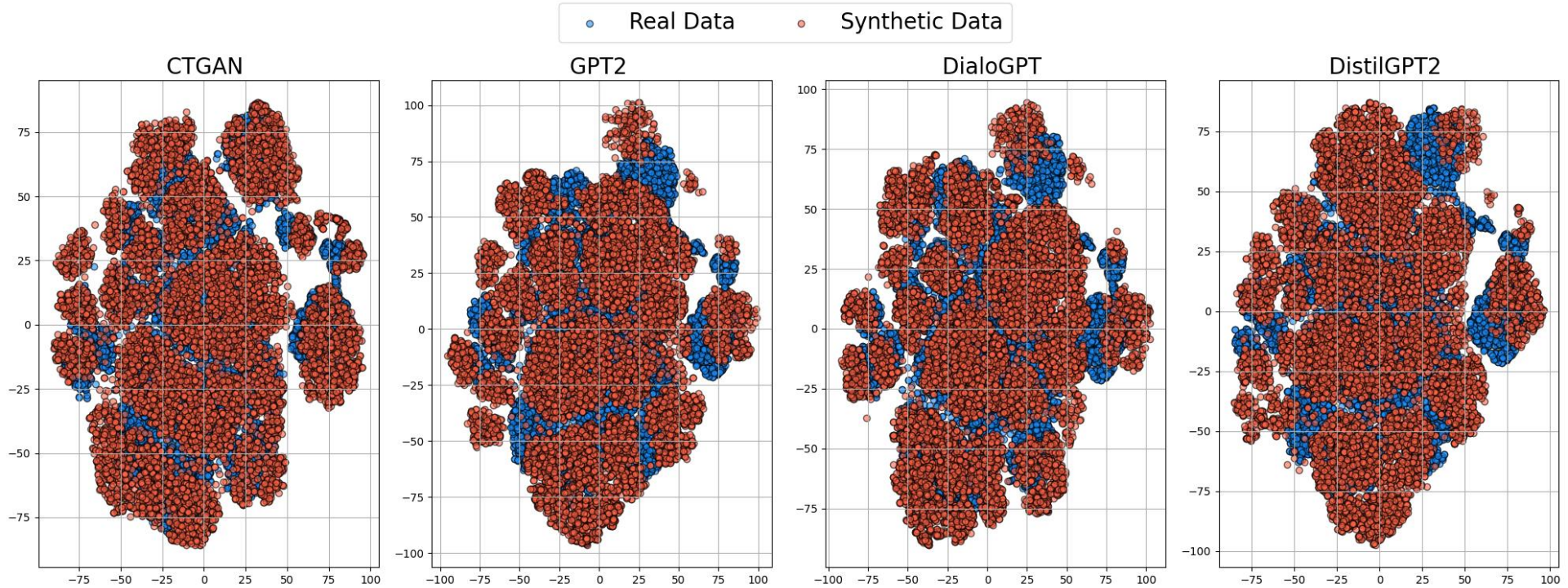


Figure 3: This figure shows the t-SNE projection comparing real and synthetic data generated by different methods for the student info 30% dataset.

Implications for Educational Research

Open science at scale

Publish privacy-safe synthetic versions of confidential educational datasets

Cross-border collaboration

Share data across jurisdictions without violating GDPR or local regulations

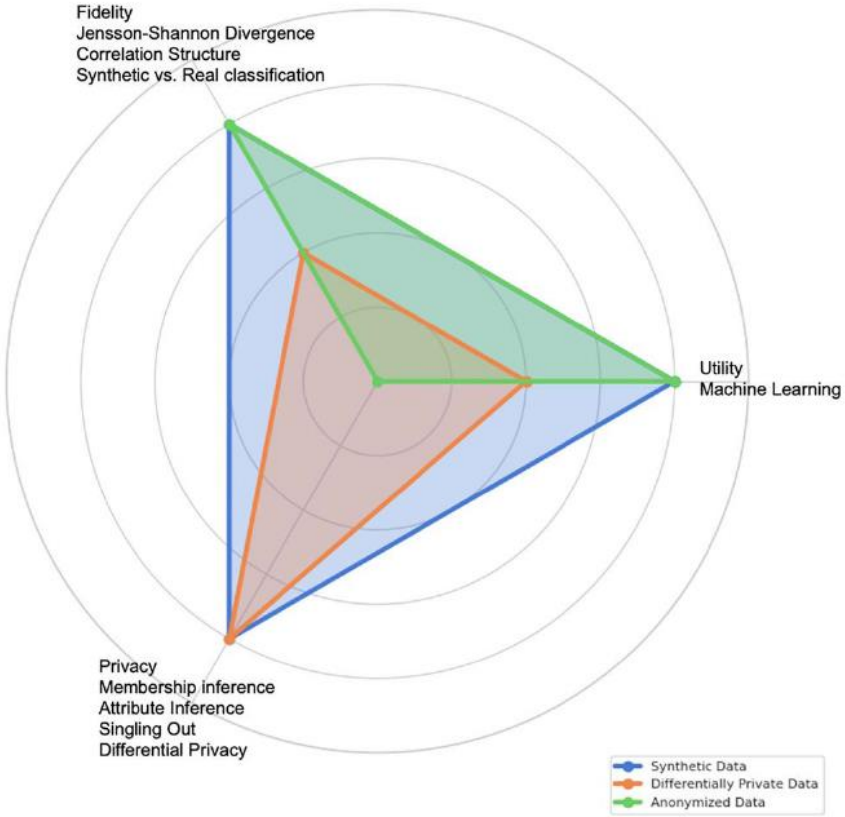
Data augmentation

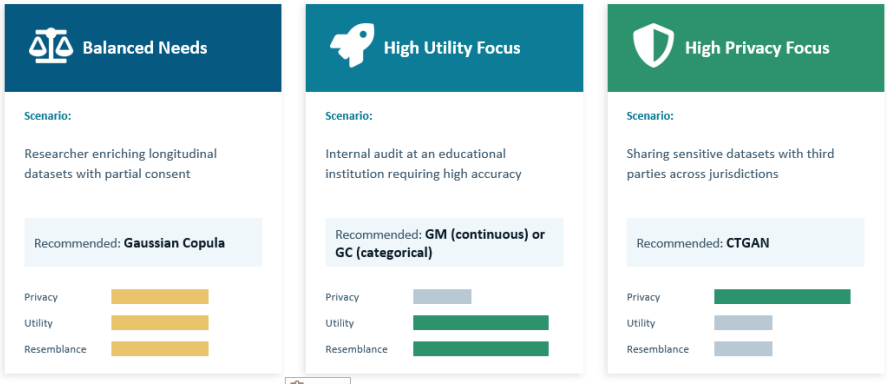
Address class imbalance and data scarcity for building robust LA models

Reproducible research

Enable replication studies with synthetic benchmarks when real data cannot be shared

The interesting trade-off in evaluating synthetic data generation





Liu, Q., Khalil, M., Jovanovic, J., & Shakya, R. (2024). *Scaling While Privacy Preserving. LAK'24.*

Adams, T., Birkenbihl, C., Otte, K., Ng, H. G., Rieling, J. A., Näher, A. F., ... & Fröhlich, H. (2025). On fidelity versus privacy and utility trade-off of synthetic patient data. *iScience*, *28*(5).

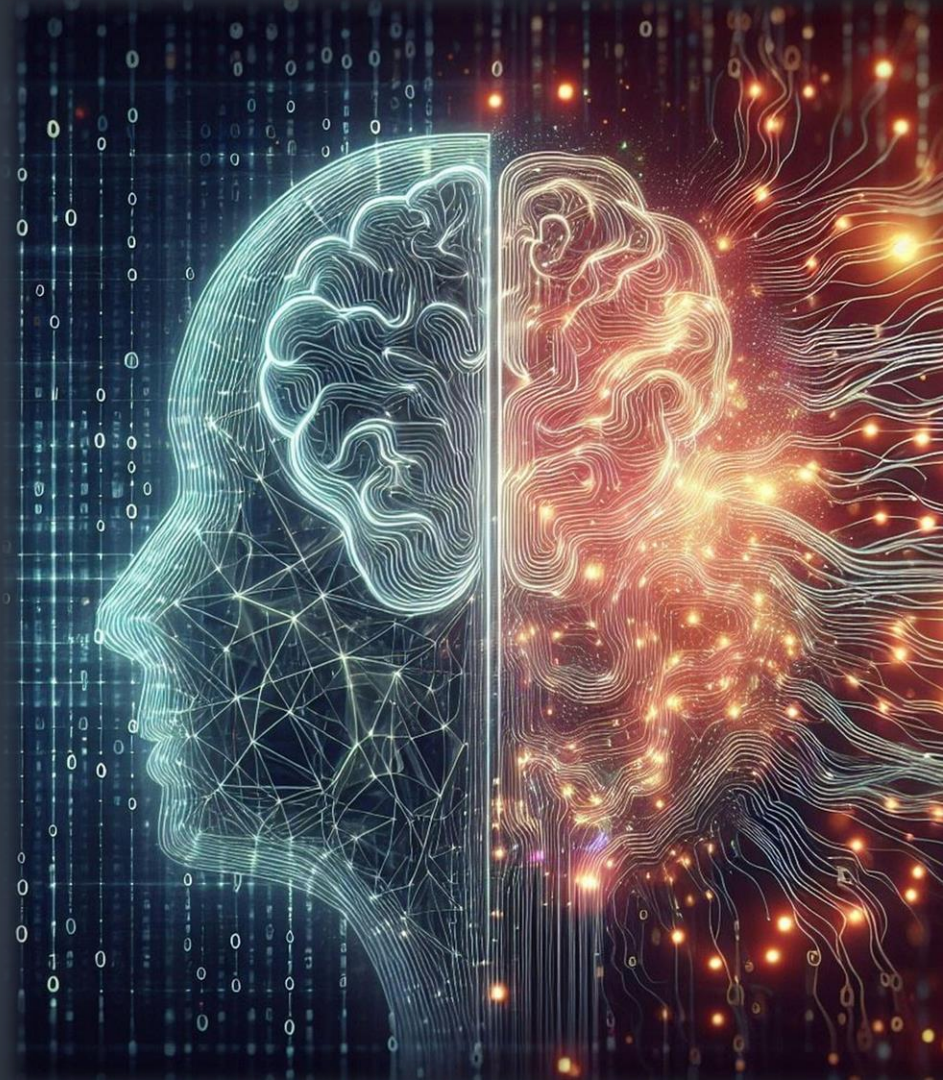
Limitations

Challenge:

Heavy Computational Demands

Challenge:

Synthetic-trained models Collapse



Challenge:

Missing Nuances

Thank You

Mohammad Khalil

mohammad.khalil@uib.no

OECD | March 2026 | Online