Generative AI for Digital Marketing Data

TEAM G:

STEPHANIE LU, **ASHWIN** RAMASESHAN,
HOCHAN SON, **JUYI** YANG

Agenda

01 Introduction

02 Model Overviews

03 Results

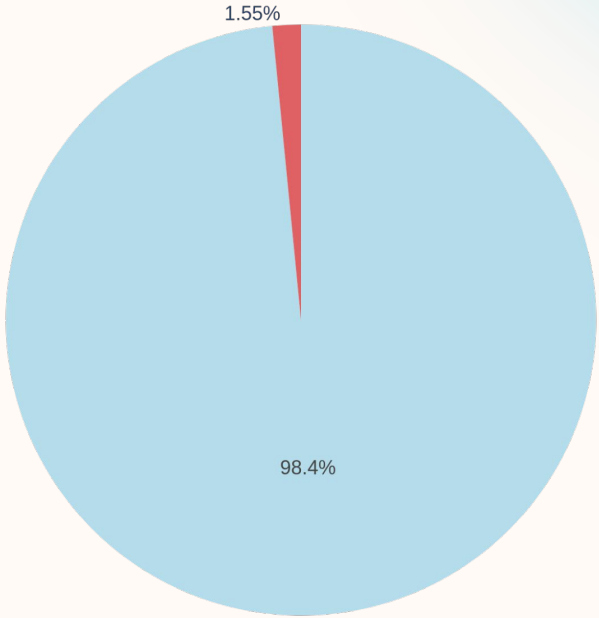
04 Implications

05 Next Steps

Introduction

- **Synthetic data** is artificially generated data that mimics real-world datasets while preserving key statistical properties
- This generative AI method helps
  - Address data **imbalance** (e.g. 98% vs. 2%)
  - Enhances model **generalization and privacy**

- Our dataset represents user engagement (CTR, demographics, device info)
- **Generative AI** (CTGAN, VAE, LLMs) helps
  - Simulate realistic user behavior
  - Improve predictive modeling for better ad targeting & campaign optimization



Model Overviews

Method	Best For
TabDDPM	Production-quality synthetic data with high fidelity
CTGAN	Quick implementation with good quality results
TVAE	When generation speed is critical
LLMs	When you need creative variations or have text-heavy features

TabDDPM

- A generative model specifically designed for **tabular data**, utilizing diffusion processes to iteratively transform noise into realistic synthetic data.
- **Dataset**: Trained on **10% of the real dataset** to create synthetic samples.

TabDDPM

Why TabDDPM?

- **Excels in modeling complex feature interactions** and distributions in mixed-type (continuous and categorical) data.
- Demonstrates **superior performance in data privacy**, preserving detailed feature-level statistical properties.

Strengths & Considerations

- Provides **high-quality synthetic data**, capturing intricate dependencies effectively.
- **Computationally intensive** due to iterative denoising processes.
- Requires careful tuning to **prevent overfitting and ensure stability** during training.

TVAE

TVAE (Tabular Variational Autoencoder)

- A generative model designed for **tabular data**, using deep learning to encode and decode structured information.
- **Dataset**: Trained on **10% of the real dataset** to generate synthetic samples.

Performance & Results

- **Loss decreased over training epochs**, indicating effective learning.
- **Maintains probability distribution similarity** with real data (JS Divergence = 0.223).

TVAE

Why TVAE?

- **More effective for complex tabular data** that has both continuous and categorical variables.
- **Can capture intricate feature dependencies**, making it useful for high-dimensional datasets.
- **Stronger performance in data privacy** while preserving data utility compared to simpler generative models.

Strengths & Considerations

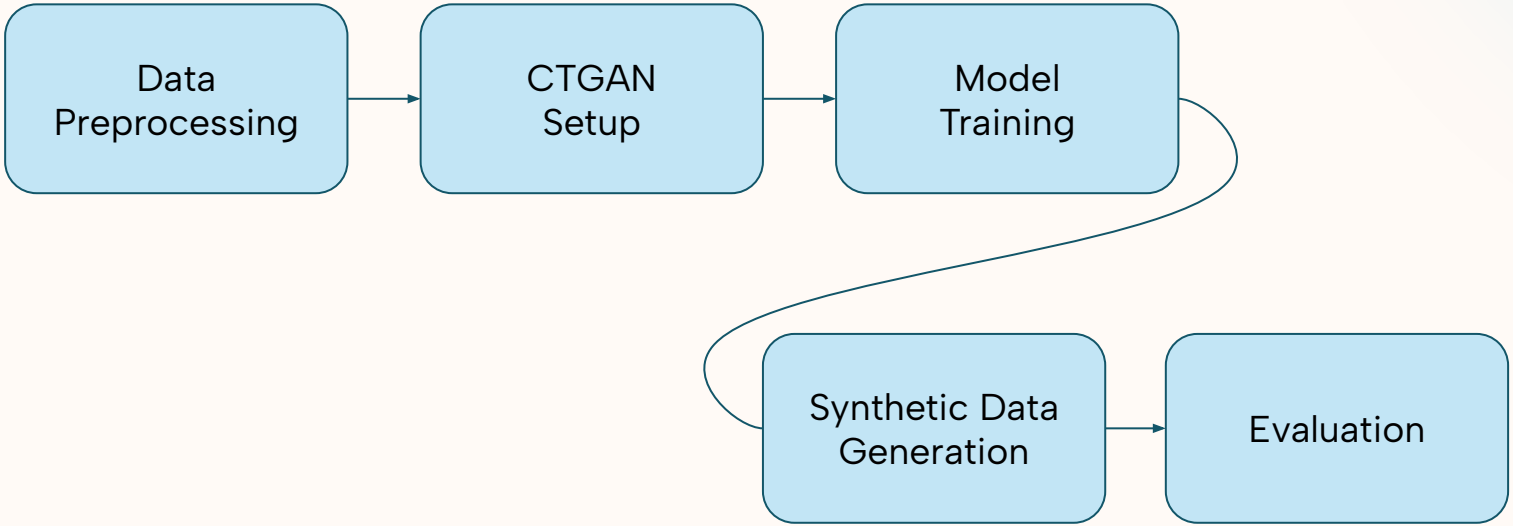
- **Captures feature relationships well**, making it suitable for structured datasets.
- **More computationally expensive** than other methods like CTGAN.
- **Potential risk of overfitting** if not properly regularized.

CTGAN

- Conditional Tabular Generative Adversarial Network
- A generative model designed to create **realistic** synthetic data
- Part of the Generative Adversarial Network (GAN) family, where two networks (the generator and the discriminator) are trained **simultaneously**

- **How CTGAN Works:**
  - Generator creates synthetic data.
  - Discriminator distinguishes between real and synthetic data.
  - Goal: for the generator to fool the discriminator by generating realistic data.
- **Why CTGAN?**
  - Ideal for generating high-dimensional and mixed-type data (categorical + numerical).
  - More stable than traditional GANs, making it effective for generating tabular data.

CTGAN


Data Preprocessing
CTGAN Setup
Model Training
Synthetic Data Generation
Evaluation

CTGAN

Weaknesses:

- Handling of high-cardinality categorical data
- Class imbalance may still persist in generated data
- Mode collapse risks result in a lack of diversity
- Overfitting can lead to poor generalization
- Training dynamics are tricky and require tuning
- High computational cost

LLMs (Large Language Models)

Base model	**GPT2**	**Deepseek-coder (1.3b)**
Developer	OpenAI (USA, 2019)	High-Flyer (China, 2023)
Quantization	Floating point 32 Bits	4 and 8 Bits
Memory requirement	> 10 GB VRAM	> 14 GB VRAM
Parameters	137 Million	1.3 Billions
Context Window	1024 tokens	16,384 tokens

GPT2: https://huggingface.co/openai-community/gpt2
Deepseek-coder-1.3b: https://huggingface.co/deepseek-ai/deepseek-coder-1.3b-base

Local LLMs Training Environments

- **Mac Mini**
- Apple Silicon **M2 Pro**
- CPU: **10** Core
- GPU: **16** Core
- RAM: **16** GB (unified)

Huggingface.co

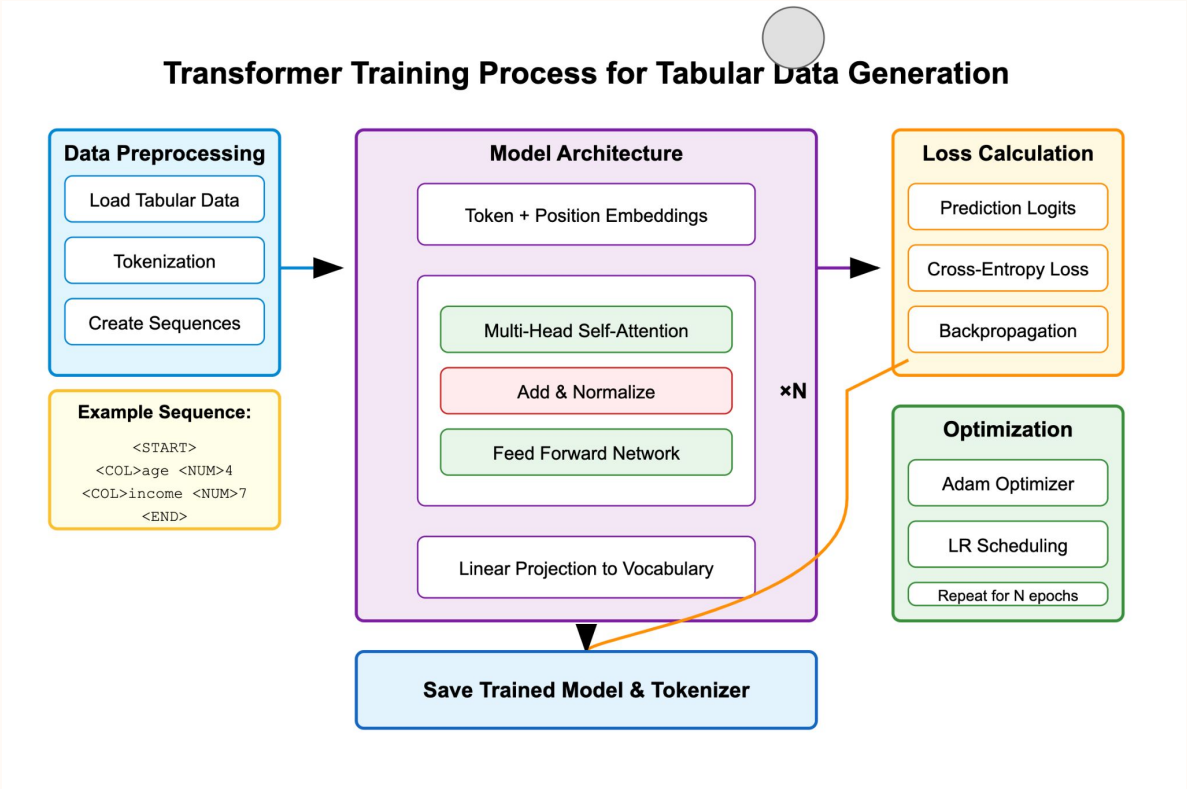
Ollama

GPT-2

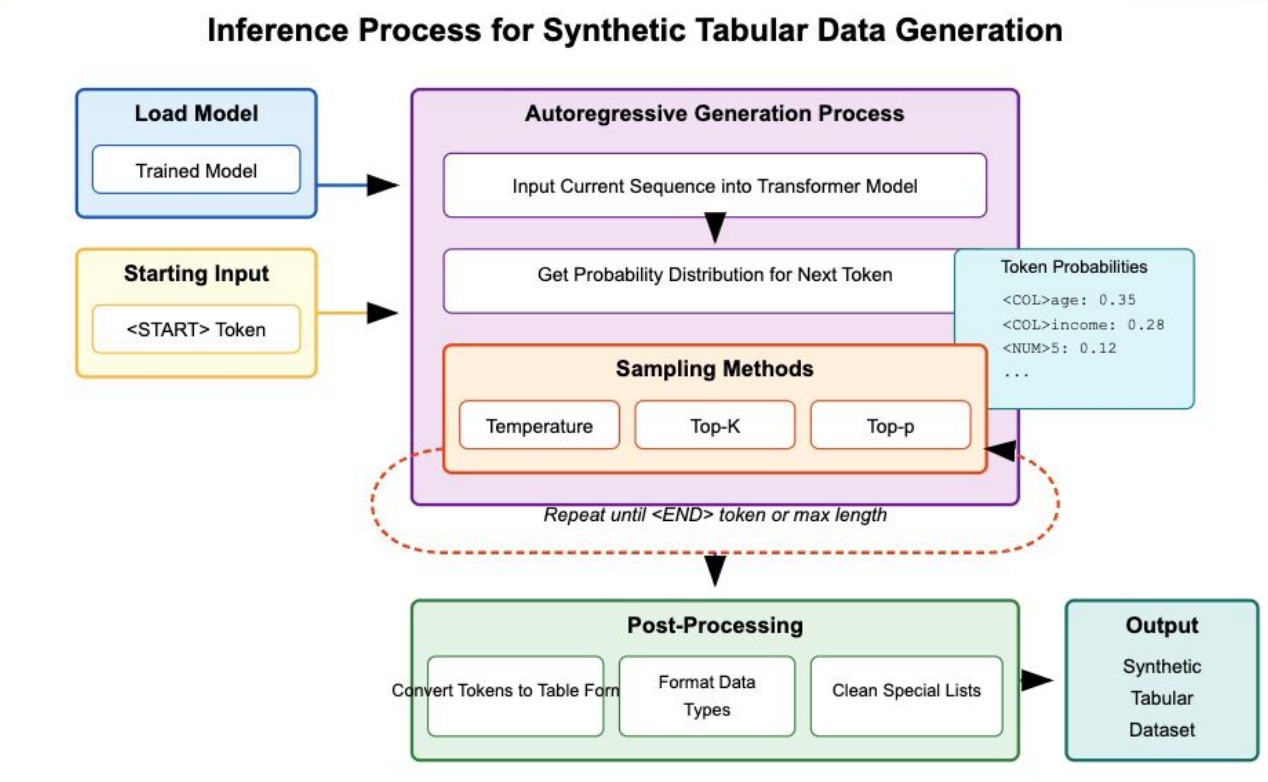
PyTorch

deepseek

Tabular LLM Training (GPT-2)

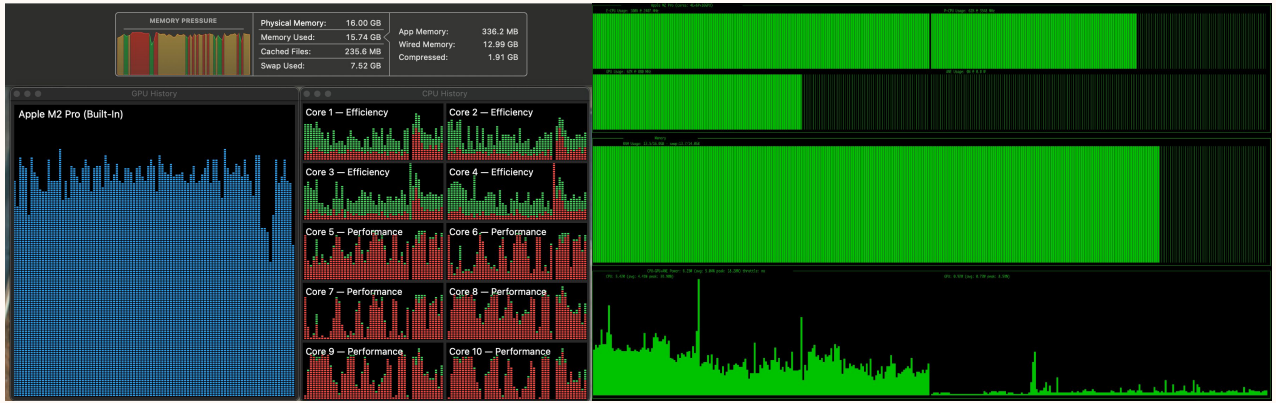


Tabular LLM Inference (GPT-2)



LLM Caveats

- **Resource**: Memory limits (16GB Unified Memory | CPU: 10 Core | GPU: 16 Core)
- **Quantization of choices**: F32, BF16, 8 bits, and 4 bits
- **GPU Hardware**: NVIDIA vs Apple Silicon
- **Python Library**: : CUDA vs MPS (Pytorch) or MLX
- **Training vs Inference performance**: 4+ hours | > 10 token/sec (1 row/1.5 min)
- **Parameter**: **Higher** is better quality, but require **more resources**
- **Older models:** (GPT-2: since 2019 vs Deepseek-coder: since 2023)



Fidelity – TVAE vs. CTGAN

Kolmogorov–Smirnov (KS) Test

- KS test compares the cumulative distributions of real and synthetic data to measure similarity.
- KS Statistic **(TVAE) = 0.758**, KS Statistic **(CTGAN) = 0.204**
- Lower KS = More fidelity (closer to real data distribution).
- **CTGAN wins** as it better matches the real data distribution.

Fidelity – TVAE vs. CTGAN

Jensen–Shannon (JS) Divergence

- JS divergence quantifies how different two probability distributions are, with lower values indicating better alignment.
- **TVAE JS = 82.45% , CTGAN JS = 4.92%**
- Lower JS = Better overlap with real data.
- **CTGAN wins** as TVAE’s divergence is too extreme.

Utility – TVAE vs. CTGAN

Mean Absolute Error & Summary

- MAE calculates the average absolute difference between real and synthetic values, assessing numeric precision.
- TVAE MAE = .265, CTGAN MAE = .195
- Lower MAE = Closer to real data values
- CTGAN excels at capturing data distribution (KS Test)
- CTGAN also has better Fidelity
- CTGAN is better for distribution matching, while TVAE is better for probability and numeric accuracy.

Model Performance Evaluation (Fidelity)

- Comparison between Baseline vs Models

Metric	Baseline	TVAE	CTGAN	LLM-GPT2	Best
Overall Fidelity (Higher is better)	100%	25.66%	94.51%	79.19%	CTGAN
KS Test (Lower is better)	0%	97.03%	6.05%	11.65%	CTGAN
JS Divergence (Lower is better)	0%	82.45%	4.92%	29.96%	CTGAN

- Comparison between Models

	TVAE vs CTGAN	TVAE vs LLM	CTGAN VS LLM	Similarity
KS Test (Lower is better)	75.8%	97.69%	6.95%	CTGAN VS LLM
JS Divergence (Lower is better)	22.3%	82.18%	5.63%	CTGAN VS LLM

Model Performance Evaluation (Utility)

Metrics	Baseline	TVAE	CTGAN	LLM-GPT2	BEST
MAE (**Lower** is better)	**0.267**	**0.265**	**0.195**	**0.209**	**CTGAN**
Accuracy (**Higher** is better)	**0.733**	**0.735**	**0.805**	**0.791**	**CTGAN**
Precision (**Higher** is better)	**0.596**	**0.723**	**0.898**	**0.884**	**CTGAN**
Recall (**Higher** is better)	**0.225**	**0.749**	**0.662**	**0.694**	**TVAE**
F1-Score (**Higher** is better)	**0.327**	**0.736**	**0.762**	**0.777**	**LLM**

**we have used XGBoost model to fit and test this scores.

Limitations

Method	Weaknesses
CTGAN	• **Mode collapse** due to adversarial training instability • Struggles with **high-cardinality categorical variables** • **Class imbalance** issues (minority class generation) • Difficult to generate **complex, highly dependent features** (e.g., sequential behaviors in digital marketing)
TVAE	• Can be **computationally expensive** during training • May struggle with **extremely high-dimensional datasets** or datasets with **very sparse categorical variables**
LLMs	• Not suited for **tabular data** due to numerical LLM data generation **high computational cost** • Requires **significant pretraining and fine-tuning** for tabular data generation • Can be **resource-intensive** and **overkill** for simple tabular datasets
TabDDPM	• **Computationally expensive** and **time-consuming** due to iterative training process • Requires **good hyperparameter tuning** to generate realistic data

Implications

- **Privacy**– Synthetic data allows for sharing and analysis without exposing sensitive user information, reducing privacy risks.
- **Enhanced Model Training** – Helps overcome data scarcity by augmenting real datasets, improving machine learning model performance.
- **Bias Considerations** – Synthetic data can mitigate biases in real-world datasets but also risks amplifying existing biases if not carefully generated.
- **Real-World Applications** – Used in healthcare, finance, and marketing to create realistic datasets for research, testing, and predictive modeling without regulatory concerns.

Next Steps

- Refine generative models via **hyperparameter optimization** (e.g., Bayesian optimization) for enhanced synthetic data quality
- Explore **advanced** generative methods (e.g. other diffusion models)
- Conduct statistical analyses to measure and reduce **bias**, ensuring synthetic data fairly represents the real dataset's distributions
- Test the effectiveness of **hybrid** training sets (real + synthetic data) to address class imbalance
- Evaluate synthetic data's **impact** on downstream tasks such as CTR prediction
- Measure the Differential Privacy

Thank You!
Questions?