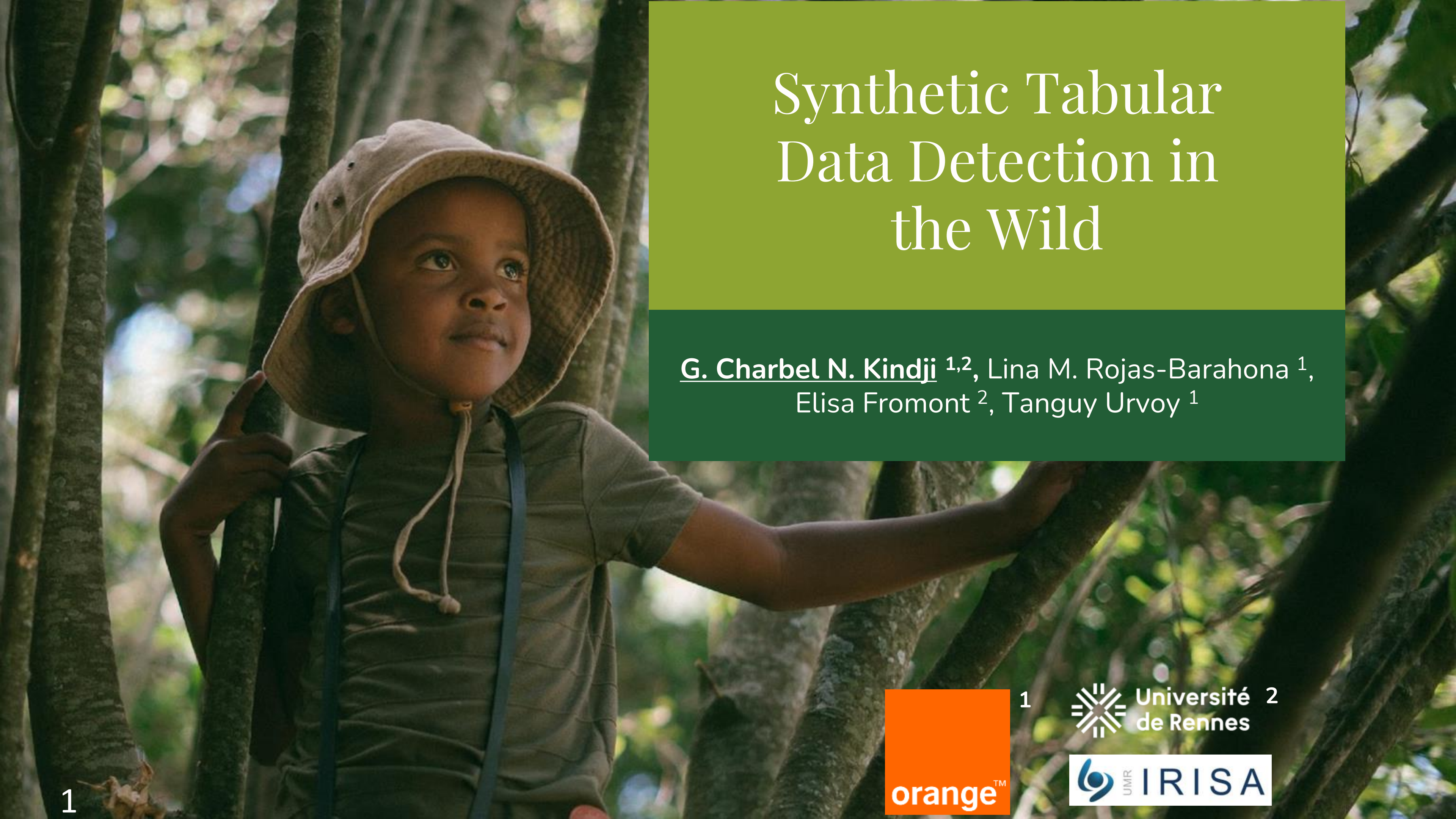


Synthetic Tabular Data Detection in the Wild

G. Charbel N. Kindji [1,2], Lina M. Rojas-Barahona [1], Elisa Fromont [2], Tanguy Urvoy [1]

orange™ [1]

Université de Rennes [2]

IRISA

Agenda

1. Introduction
2. Detection in the Wild
3. Experimental Setup
4. Experimental Results
5. Final Remarks

Introduction (1/3)

Misuse of generative models

Credit: YouTube French Faker

Credit : David Fathi / Midjourney

More effective generative models

Text, image, audio, video

Risk:
Data forgery

Eg. Fake images and videos

It is important to develop detectors

Detecting synthetic data

Challenge in real life scenario

Detection → Binary classification

Detectors struggle with new content (in the ''wildness'' of real life)

Introduction (2/3)

Detectors struggle with new content

(« in the wildness of real life »)

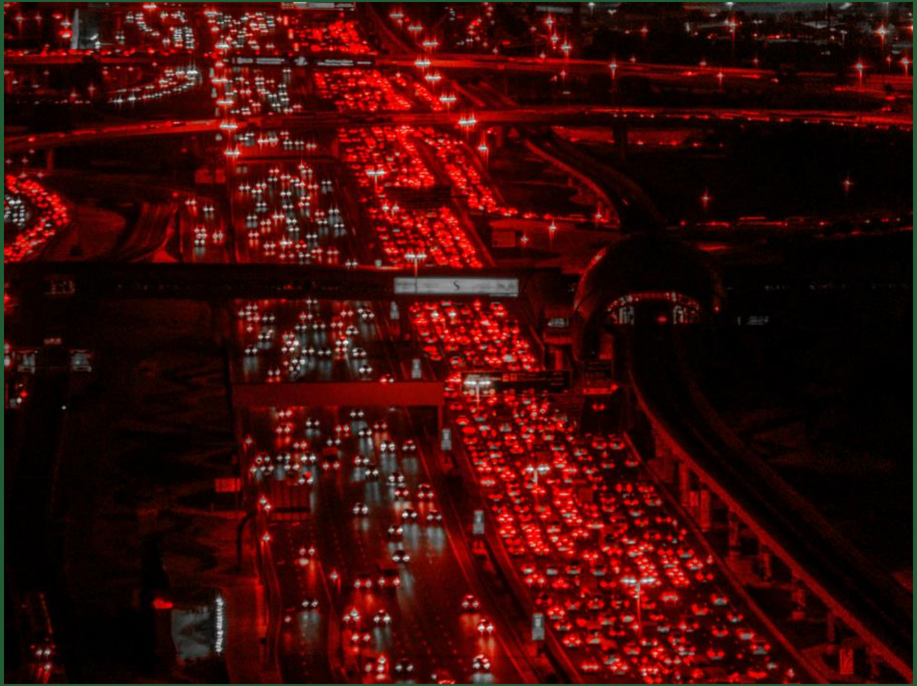
Controlled Environment

Introduction (2/3)

Detectors struggle with new content

(« in the wildness of real life »)

Controlled Environment **vs.** **In the Wildness of Real Life**



Introduction (2/3)

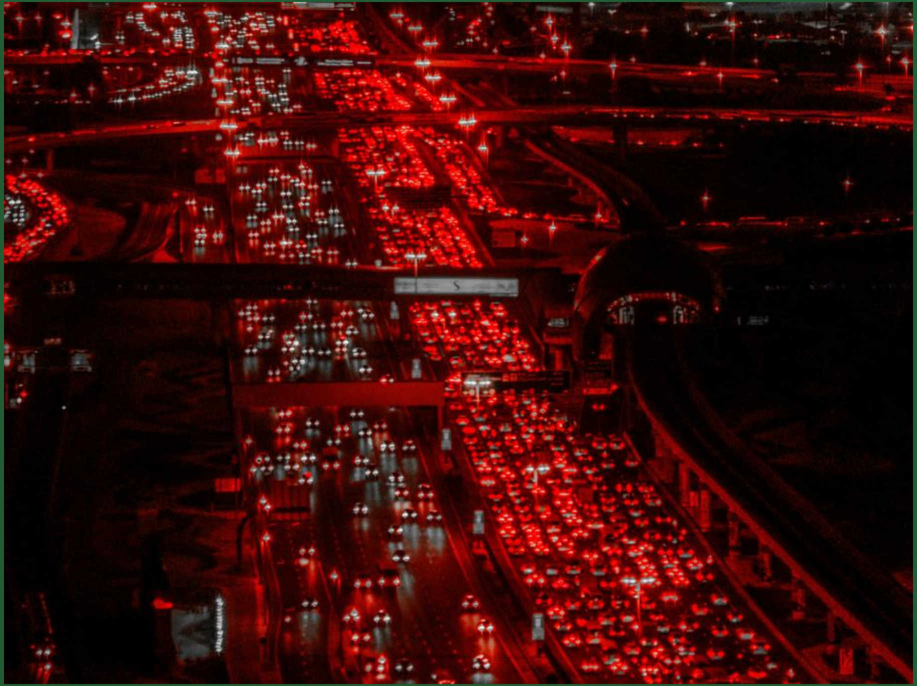
Detectors struggle with new content

(« in the wildness of real life »)

Eg.:

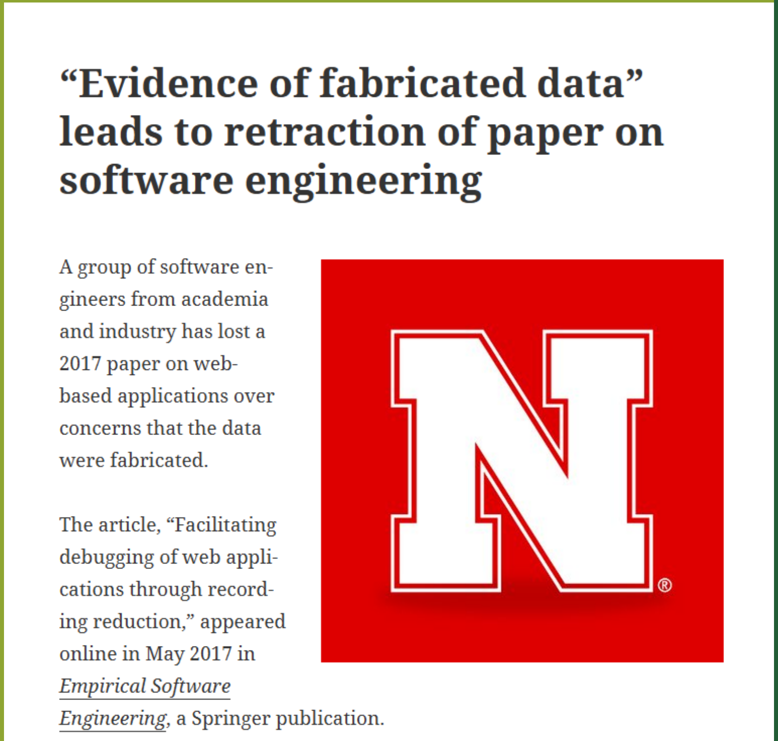
- Unknown generators
- Domain shift
- Adversarial setting

Controlled Environment vs. In the Wildness of Real Life



Introduction (3/3)

Misuse of generative models: Focus on Tabular Data



“Evidence of fabricated data” leads to retraction of paper on software engineering

A group of software engineers from academia and industry has lost a 2017 paper on web-based applications over concerns that the data were fabricated.

The article, “Facilitating debugging of web applications through recording reduction,” appeared online in May 2017 in *Empirical Software Engineering*, a Springer publication.

Source: https://retractionwatch.com/2019/01/24/evidence-of-fabricated-data-leads-to-retraction-of-paper-on-software-engineering/

Tabular Data Generation → Hot Topic

General and domain-specific tabular data generators

High Quality Tabular Data Generators

TabDDPM [1], TabSyn [2]

Data Forgery

Eg.:

- Fake accounting tables
- Fake scientific results

Specific Table Issue: Cross-table Shift

Change in the table structure at detector’s deployment

[1] Kotelnikov et al., TabDDPM: Modelling Tabular Data with Diffusion Models, ICML 2023
[2] Zhang et al., Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space, ICLR 2024

Detection in the Wild

Focus of our study

- Synthetic Tabular Data Detection → Classification problem
- Can be done on the same table structure
- Classifier Two Sample Test Metric [1,2]

Product ID	Price	Rating	Label
P001	19.99	4.5	Real
P265	29.99	3.0	Real
P4565	199.99	5.0	Synthetic
P018	39.99	4.2	Real
P107	100.00	8.5	Synthetic

Detection in the Wild

Focus of our study

[1] Lopez-Paz, D., Oquab, M.: Revisiting classifier two-sample tests. ICLR 2016
[2] G. Charbel N. Kindji, Lina Maria Rojas-Barahona, Elisa Fromont, Tanguy Urvoy. Under the Hood of Tabular Data Generation Models: Benchmarks with Extensive Tuning. 2024.

- Synthetic Tabular Data Detection → Classification problem
- Can be done on the same table structure
- Classifier Two Sample Test Metric [1,2]

Product ID	Price	Rating	Label
P001	19.99	4.5	Real
P265	29.99	3.0	Real
P4565	199.99	5.0	Synthetic
P018	39.99	4.2	Real
P107	100.00	8.5	Synthetic

Detection in the Wild

Focus of our study

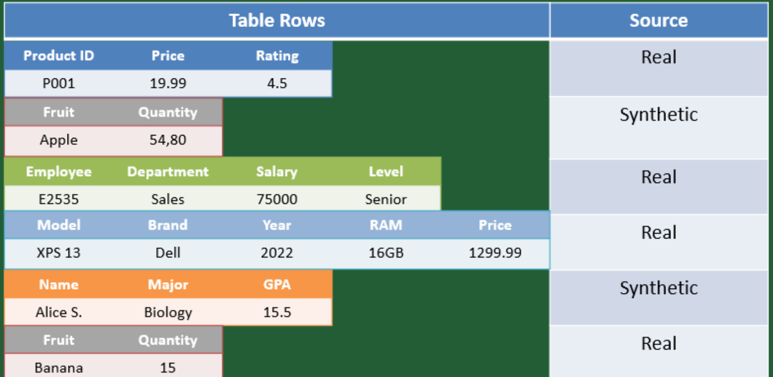
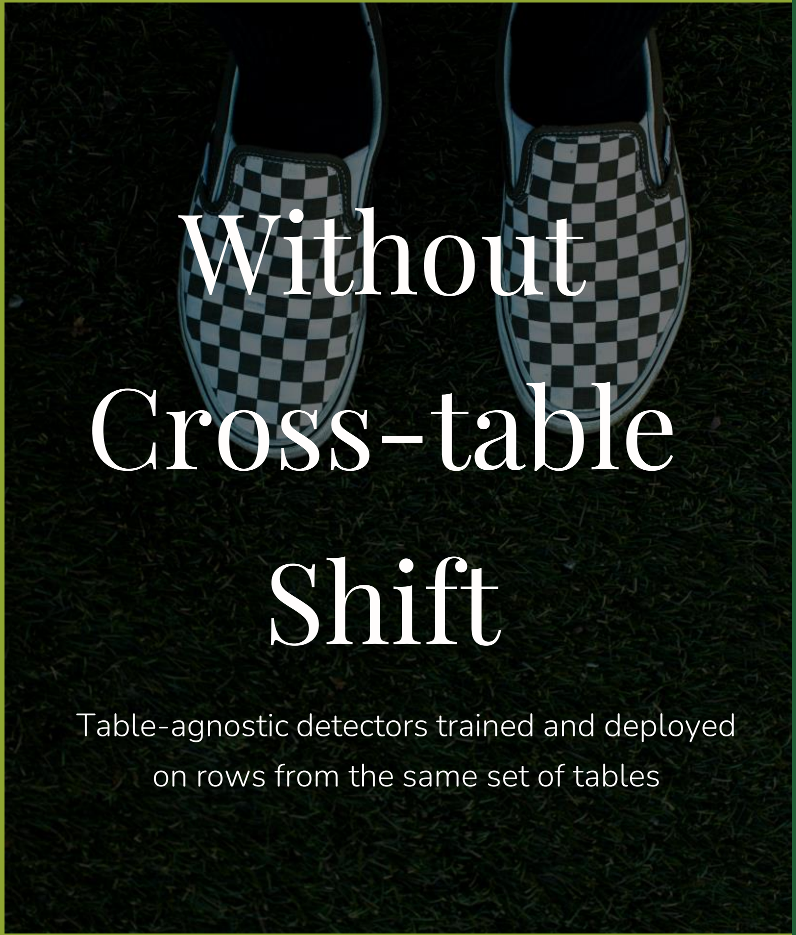


Table Rows					Source
Product ID	Price	Rating			Real
P001	19.99	4.5			
Fruit	Quantity				Synthetic
Apple	54,80				
Employee	Department	Salary	Level		Real
E2535	Sales	75000	Senior		
Model	Brand	Year	RAM	Price	Real
XPS 13	Dell	2022	16GB	1299.99	
Name	Major	GPA			Synthetic
Alice S.	Biology	15.5			
Fruit	Quantity				Real
Banana	15				

- Requires table-agnostic detectors
- Different levels of ''wildness'': With and Without cross-table shift

[1] Lopez-Paz, D., Oquab, M.: Revisiting classifier two-sample tests. ICLR 2016
[2] G. Charbel N. Kindji, Lina Maria Rojas-Barahona, Elisa Fromont, Tanguy Urvoy. Under the Hood of Tabular Data Generation Models: Benchmarks with Extensive Tuning. 2024.



Without Cross-table Shift

Table-agnostic detectors trained and deployed on rows from the same set of tables

Example Rows from Train Tables

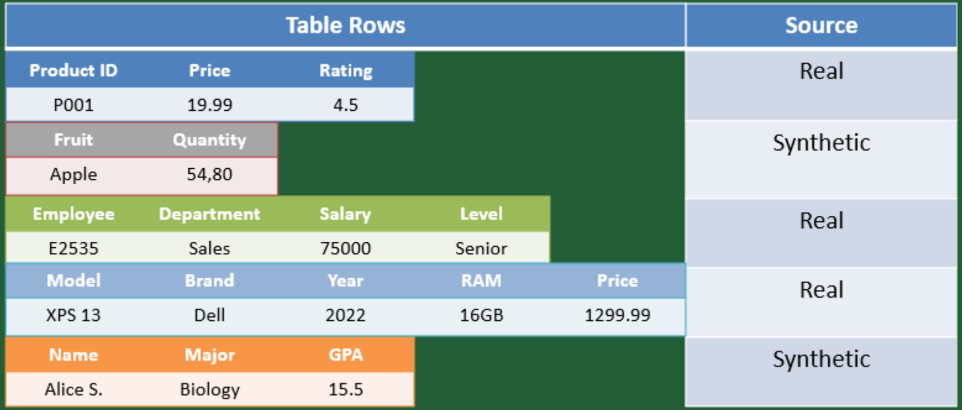


Table Rows					Source
Product ID	Price	Rating			Real
P001	19.99	4.5			
Fruit	Quantity				Synthetic
Apple	54,80				
Employee	Department	Salary	Level		Real
E2535	Sales	75000	Senior		
Model	Brand	Year	RAM	Price	Real
XPS 13	Dell	2022	16GB	1299.99	
Name	Major	GPA			Synthetic
Alice S.	Biology	15.5			

Example Rows from Test Tables

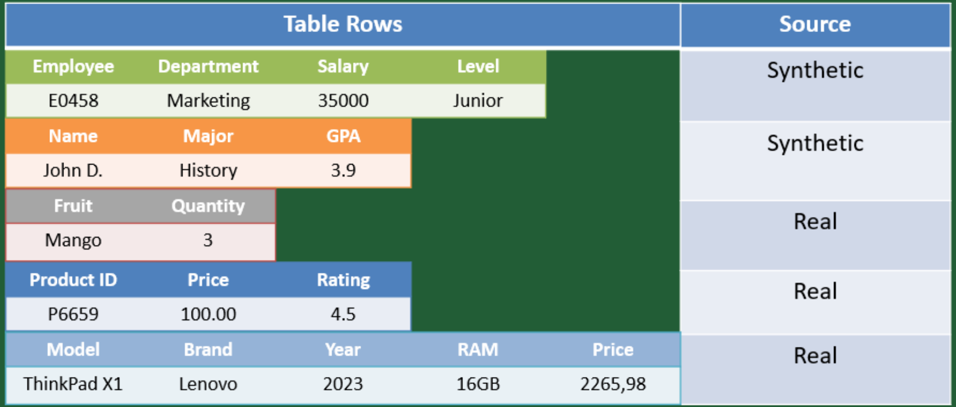


Table Rows					Source
Employee	Department	Salary	Level		Synthetic
E0458	Marketing	35000	Junior		
Name	Major	GPA			Synthetic
John D.	History	3.9			
Fruit	Quantity				Real
Mango	3				
Product ID	Price	Rating			Real
P6659	100.00	4.5			
Model	Brand	Year	RAM	Price	Real
ThinkPad X1	Lenovo	2023	16GB	2265,98	

With Cross-table Shift

Table-agnostic detectors trained and deployed on distinct set of tables

Example Rows from Train Tables

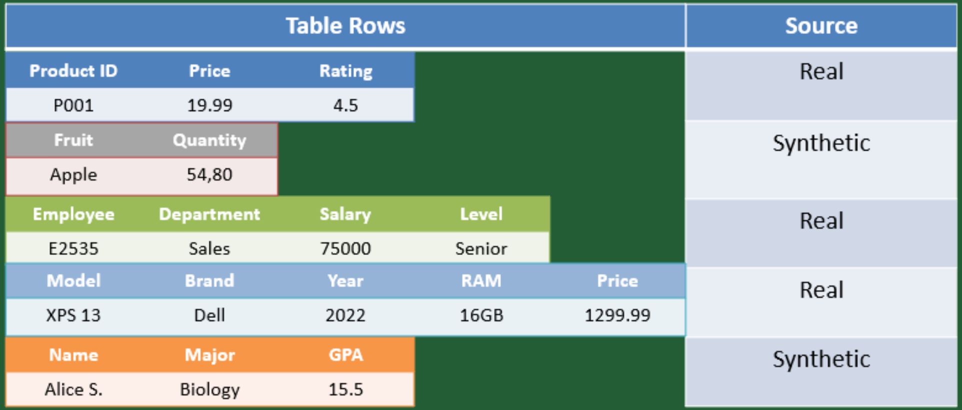


Table Rows					Source
Product ID	Price	Rating			Real
P001	19.99	4.5			
Fruit	Quantity				Synthetic
Apple	54,80				
Employee	Department	Salary	Level		Real
E2535	Sales	75000	Senior		
Model	Brand	Year	RAM	Price	Real
XPS 13	Dell	2022	16GB	1299.99	
Name	Major	GPA			Synthetic
Alice S.	Biology	15.5			

Example Rows from Test Tables

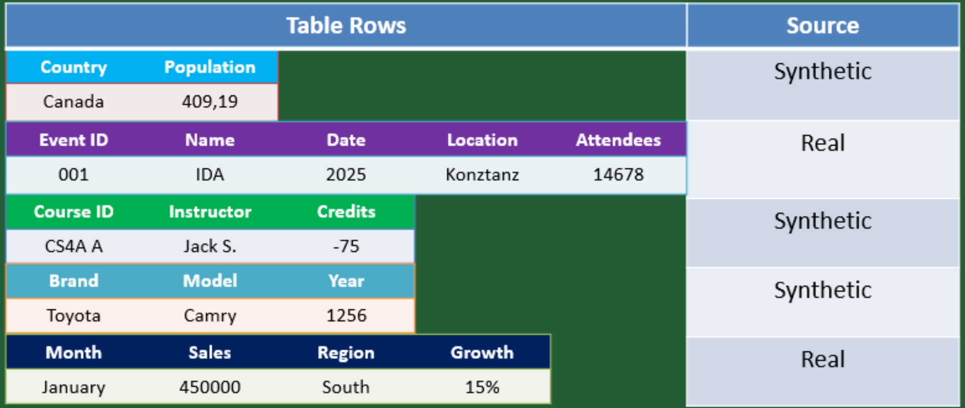


Table Rows					Source
Country	Population				Synthetic
Canada	409,19				
Event ID	Name	Date	Location	Attendees	Real
001	IDA	2025	Konztanz	14678	
Course ID	Instructor	Credits			Synthetic
CS4A A	Jack S.	-75			
Brand	Model	Year			Synthetic
Toyota	Camry	1256			
Month	Sales	Region	Growth		Real
January	450000	South	15%		

Table-agnostic encodings

Text and column-based encodings

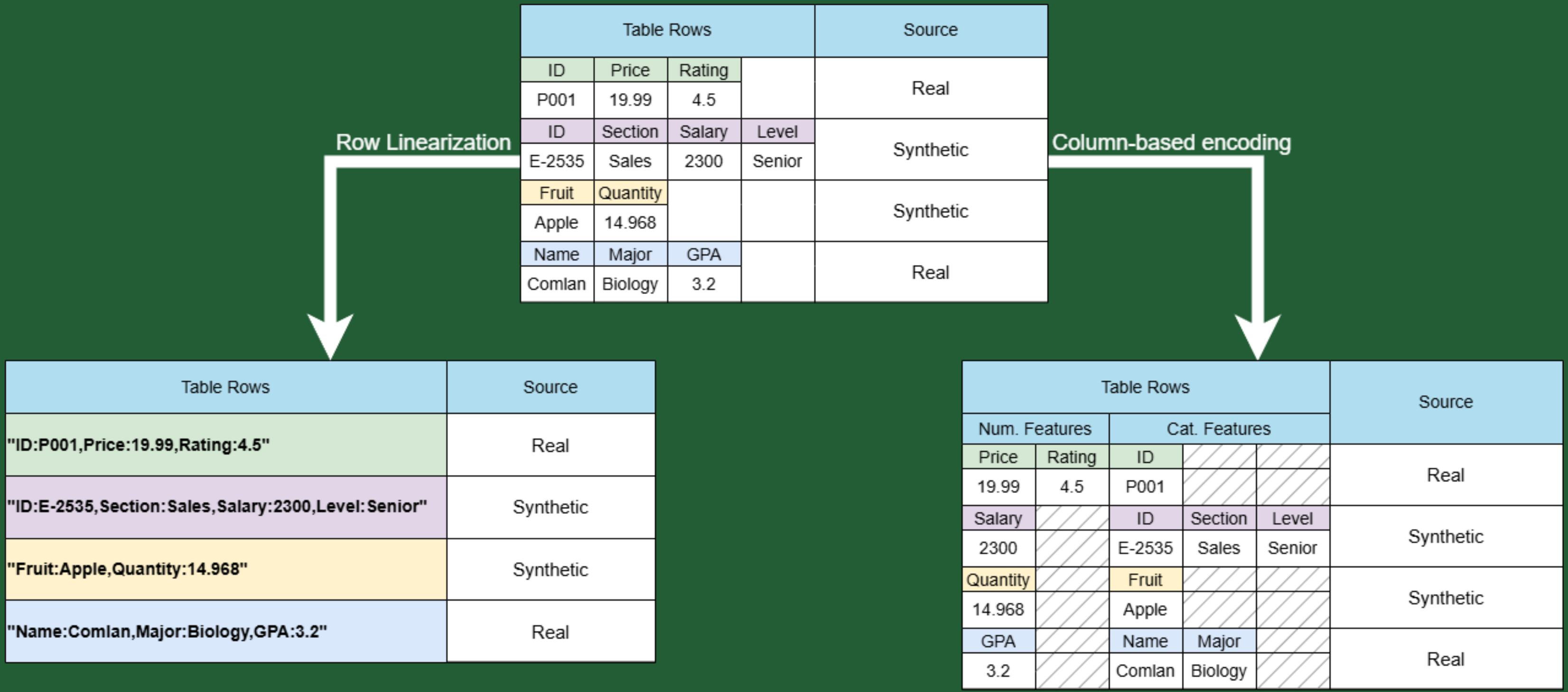


Table Rows				Source
ID	Price	Rating		Real
P001	19.99	4.5		
ID	Section	Salary	Level	Synthetic
E-2535	Sales	2300	Senior	
Fruit	Quantity			Synthetic
Apple	14.968			
Name	Major	GPA		Real
Comlan	Biology	3.2		

Row Linearization

Column-based encoding

Table Rows	Source
"ID:P001,Price:19.99,Rating:4.5"	Real
"ID:E-2535,Section:Sales,Salary:2300,Level:Senior"	Synthetic
"Fruit:Apple,Quantity:14.968"	Synthetic
"Name:Comlan,Major:Biology,GPA:3.2"	Real

Table Rows					Source
Num. Features		Cat. Features			
Price	Rating	ID			Real
19.99	4.5	P001			
Salary		ID	Section	Level	Synthetic
2300		E-2535	Sales	Senior	
Quantity		Fruit			Synthetic
14.968		Apple			
GPA		Name	Major		Real
3.2		Comlan	Biology		

Table-agnostic encodings

Text and column-based encodings

Encodings deployed on 4 detectors: XGBoost, Logistic Regression and two transformer-based detector baselines

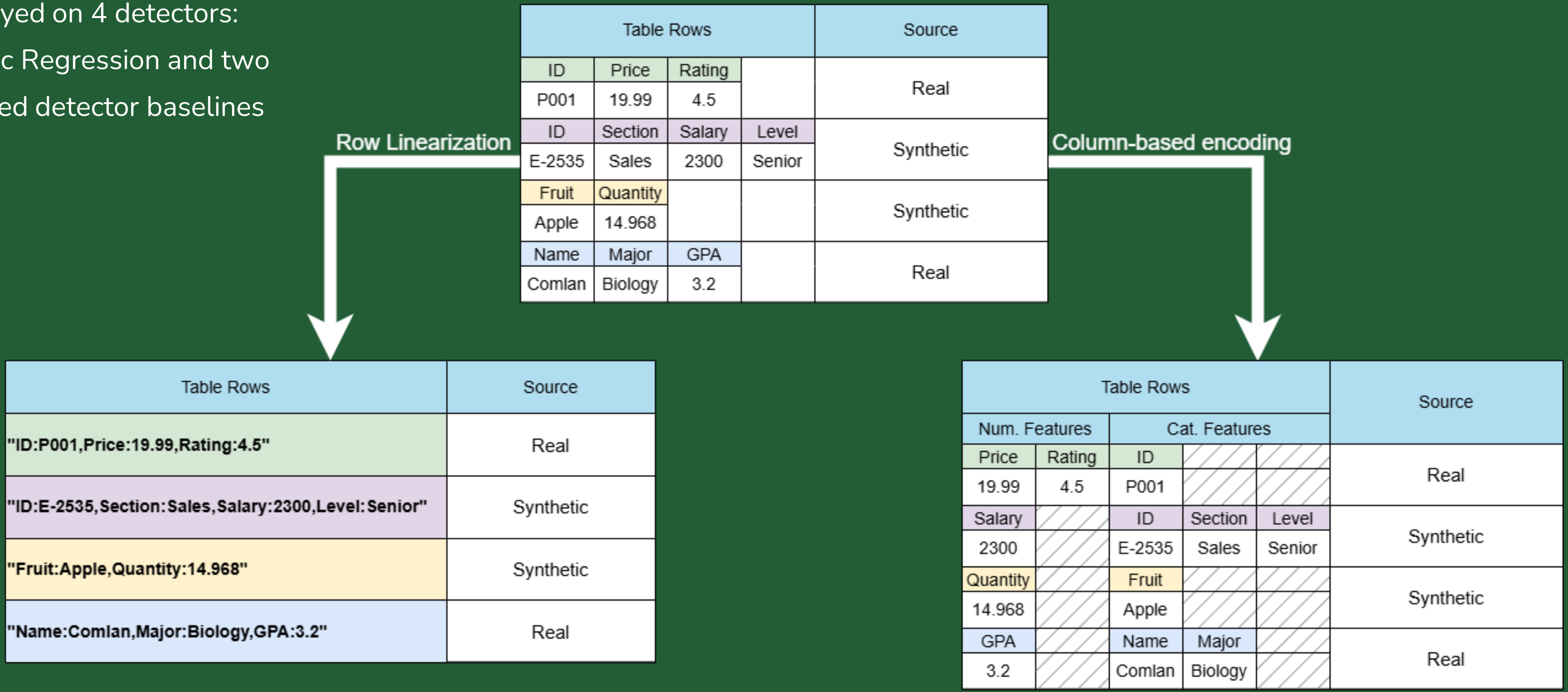


Table Rows				Source
ID	Price	Rating		Real
P001	19.99	4.5		
ID	Section	Salary	Level	Synthetic
E-2535	Sales	2300	Senior	
Fruit	Quantity			Synthetic
Apple	14.968			
Name	Major	GPA		Real
Comlan	Biology	3.2		

Table Rows	Source
"ID:P001,Price:19.99,Rating:4.5"	Real
"ID:E-2535,Section:Sales,Salary:2300,Level:Senior"	Synthetic
"Fruit:Apple,Quantity:14.968"	Synthetic
"Name:Comlan,Major:Biology,GPA:3.2"	Real

Table Rows					Source
Num. Features		Cat. Features			
Price	Rating	ID	▨	▨	Real
19.99	4.5	P001	▨	▨	
Salary	▨	ID	Section	Level	Synthetic
2300	▨	E-2535	Sales	Senior	
Quantity	▨	Fruit	▨	▨	Synthetic
14.968	▨	Apple	▨	▨	
GPA	▨	Name	Major	▨	Real
3.2	▨	Comlan	Biology	▨	

Table-agnostic encodings

Text and column-based encodings

Encodings deployed on 4 detectors: XGBoost, Logistic Regression and two transformer-based detector baselines

We also build trigrams of words and characters from linearized rows and evaluate them.

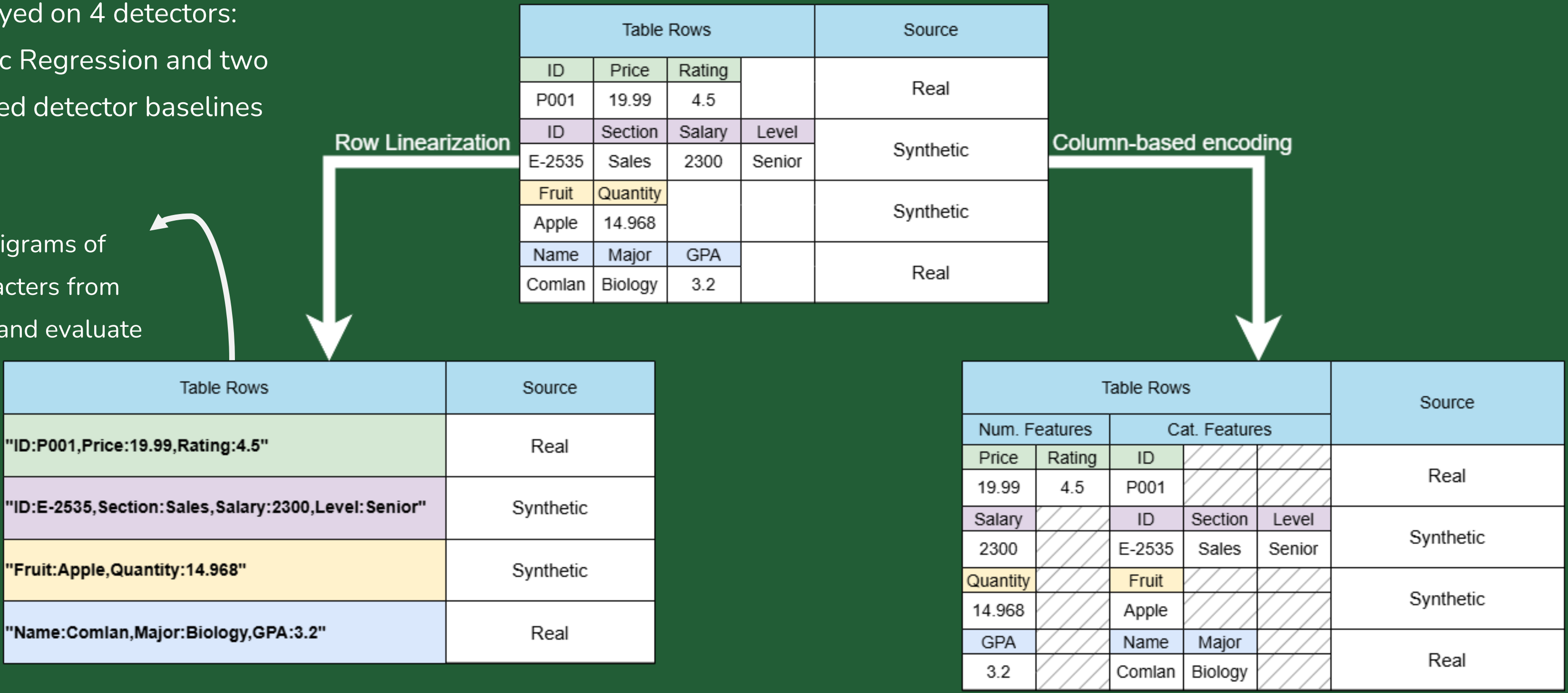
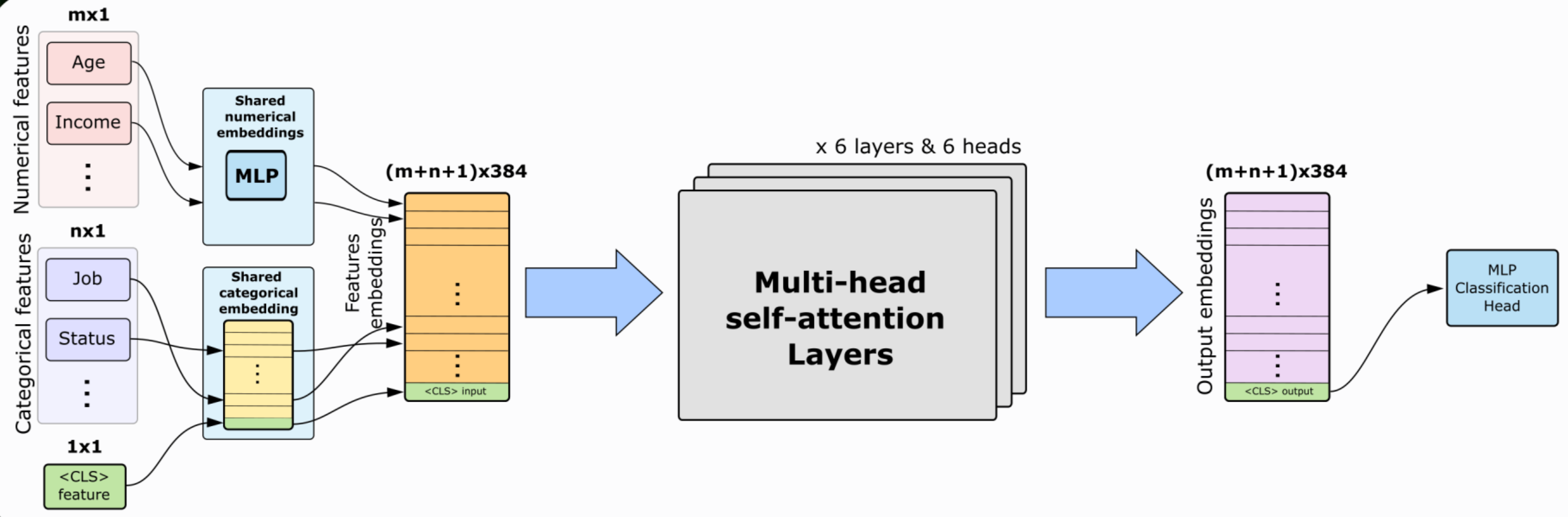


Table Rows				Source
ID	Price	Rating		Real
P001	19.99	4.5		
ID	Section	Salary	Level	Synthetic
E-2535	Sales	2300	Senior	
Fruit	Quantity			Synthetic
Apple	14.968			
Name	Major	GPA		Real
Comlan	Biology	3.2		

Table Rows	Source
"ID:P001,Price:19.99,Rating:4.5"	Real
"ID:E-2535,Section:Sales,Salary:2300,Level:Senior"	Synthetic
"Fruit:Apple,Quantity:14.968"	Synthetic
"Name:Comlan,Major:Biology,GPA:3.2"	Real

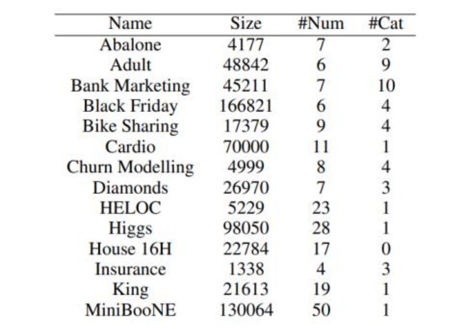
Table Rows					Source
Num. Features		Cat. Features			
Price	Rating	ID	▨	▨	Real
19.99	4.5	P001			
Salary	▨	ID	Section	Level	Synthetic
2300		E-2535	Sales	Senior	
Quantity	▨	Fruit	▨	▨	Synthetic
14.968		Apple			
GPA	▨	Name	Major	▨	Real
3.2		Comlan	Biology		

Column-based Transformer



Experimental Setup

14 Tables – UCI and Kaggle



Name	Size	#Num	#Cat
Abalone	4177	7	2
Adult	48842	6	9
Bank Marketing	45211	7	10
Black Friday	166821	6	4
Bike Sharing	17379	9	4
Cardio	70000	11	1
Churn Modelling	4999	8	4
Diamonds	26970	7	3
HELOC	5229	23	1
Higgs	98050	28	1
House 16H	22784	17	0
Insurance	1338	4	3
King	21613	19	1
MiniBooNE	130064	50	1

4 Generators

- TVAE [1]
- CTGAN [1]
- TabDDPM [2]
- TabSyn [3]

4 Detectors

- Logistic Regression
- XGBoost
- Text-Based Transformer
- Column-based Transformer

3 Setups

- Without cross-table shift: training and testing on the same set of tables
  - Single Generator
  - All Generators
- With cross-table shift: testing on a distinct set of tables

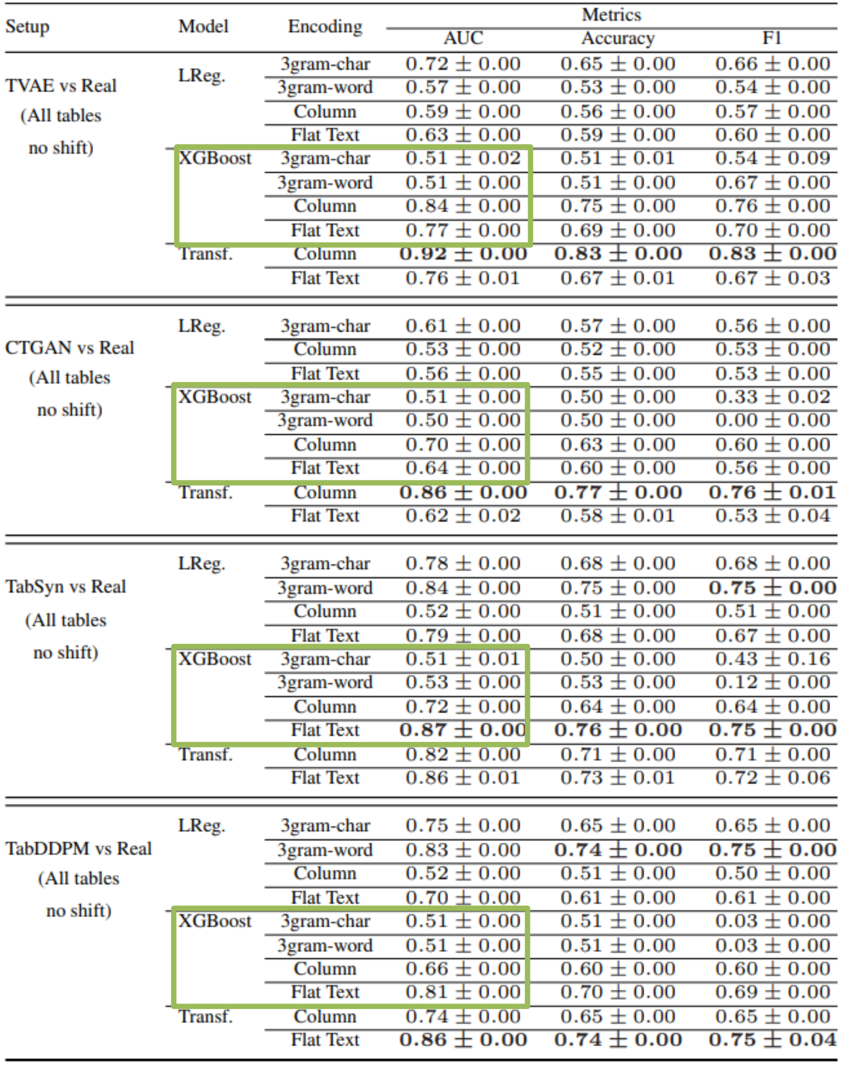
[1] Xu et al., Modeling Tabular data using Conditional GAN, NeurIPS 2019
[2] Kotelnikov et al., TabDDPM: Modelling Tabular Data with Diffusion Models, ICML 2023
[3] Zhang et al., Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space, ICLR 2024

Experimental Results – Without cross-table shift

No Table Shift Setup – Training and deploying on rows from the same set of tables

Experimental Results – Without cross-table shift (1/2)

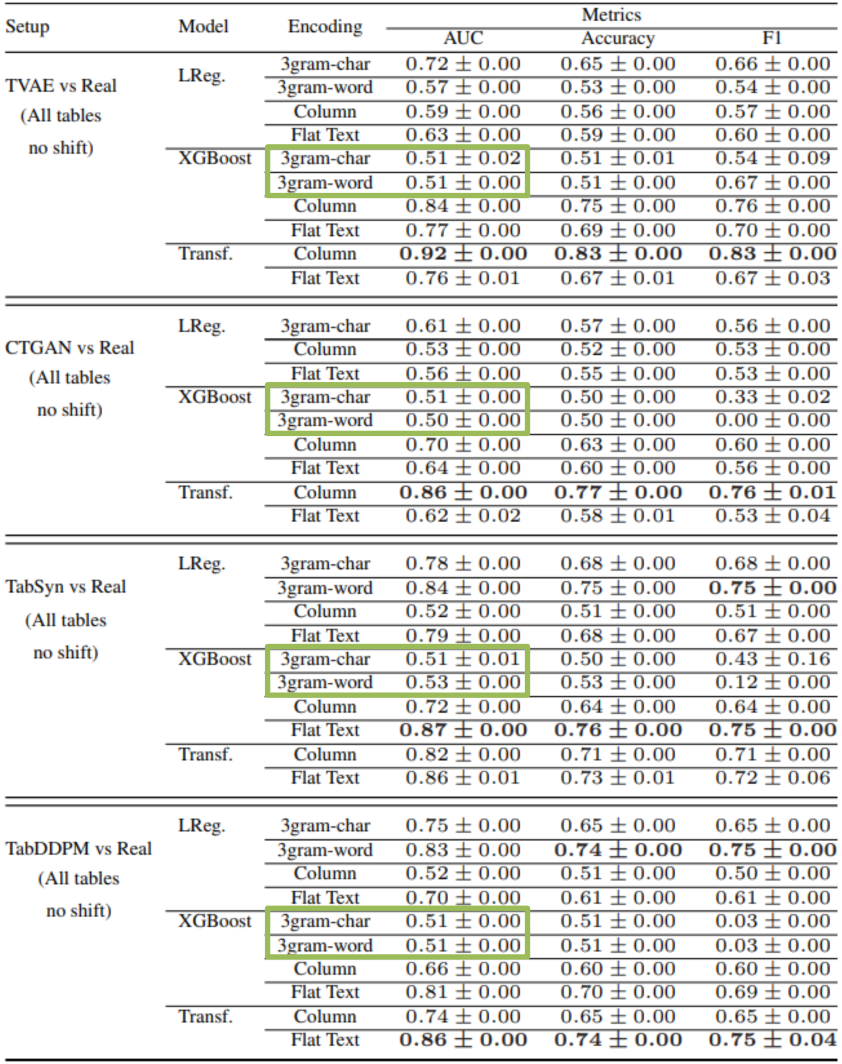
- Encoding matters, eg. XGBoost performance variation



Setup	Model	Encoding	Metrics		
			AUC	Accuracy	F1
TVAE vs Real (All tables no shift)	LReg.	3gram-char	0.72 ± 0.00	0.65 ± 0.00	0.66 ± 0.00
		3gram-word	0.57 ± 0.00	0.53 ± 0.00	0.54 ± 0.00
		Column	0.59 ± 0.00	0.56 ± 0.00	0.57 ± 0.00
		Flat Text	0.63 ± 0.00	0.59 ± 0.00	0.60 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.02	0.51 ± 0.01	0.54 ± 0.09
		3gram-word	0.51 ± 0.00	0.51 ± 0.00	0.67 ± 0.00
		Column	0.84 ± 0.00	0.75 ± 0.00	0.76 ± 0.00
		Flat Text	0.77 ± 0.00	0.69 ± 0.00	0.70 ± 0.00
	Transf.	Column	**0.92 ± 0.00**	**0.83 ± 0.00**	**0.83 ± 0.00**
		Flat Text	0.76 ± 0.01	0.67 ± 0.01	0.67 ± 0.03
CTGAN vs Real (All tables no shift)	LReg.	3gram-char	0.61 ± 0.00	0.57 ± 0.00	0.56 ± 0.00
		Column	0.53 ± 0.00	0.52 ± 0.00	0.53 ± 0.00
		Flat Text	0.56 ± 0.00	0.55 ± 0.00	0.53 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.00	0.50 ± 0.00	0.33 ± 0.02
		3gram-word	0.50 ± 0.00	0.50 ± 0.00	0.00 ± 0.00
		Column	0.70 ± 0.00	0.63 ± 0.00	0.60 ± 0.00
		Flat Text	0.64 ± 0.00	0.60 ± 0.00	0.56 ± 0.00
	Transf.	Column	**0.86 ± 0.00**	**0.77 ± 0.00**	**0.76 ± 0.01**
		Flat Text	0.62 ± 0.02	0.58 ± 0.01	0.53 ± 0.04
TabSyn vs Real (All tables no shift)	LReg.	3gram-char	0.78 ± 0.00	0.68 ± 0.00	0.68 ± 0.00
		3gram-word	0.84 ± 0.00	0.75 ± 0.00	**0.75 ± 0.00**
		Column	0.52 ± 0.00	0.51 ± 0.00	0.51 ± 0.00
		Flat Text	0.79 ± 0.00	0.68 ± 0.00	0.67 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.01	0.50 ± 0.00	0.43 ± 0.16
		3gram-word	0.53 ± 0.00	0.53 ± 0.00	0.12 ± 0.00
		Column	0.72 ± 0.00	0.64 ± 0.00	0.64 ± 0.00
		Flat Text	**0.87 ± 0.00**	**0.76 ± 0.00**	**0.75 ± 0.00**
	Transf.	Column	0.82 ± 0.00	0.71 ± 0.00	0.71 ± 0.00
		Flat Text	0.86 ± 0.01	0.73 ± 0.01	0.72 ± 0.06
TabDDPM vs Real (All tables no shift)	LReg.	3gram-char	0.75 ± 0.00	0.65 ± 0.00	0.65 ± 0.00
		3gram-word	0.83 ± 0.00	**0.74 ± 0.00**	**0.75 ± 0.00**
		Column	0.52 ± 0.00	0.51 ± 0.00	0.50 ± 0.00
		Flat Text	0.70 ± 0.00	0.61 ± 0.00	0.61 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.00	0.51 ± 0.00	0.03 ± 0.00
		3gram-word	0.51 ± 0.00	0.51 ± 0.00	0.03 ± 0.00
		Column	0.66 ± 0.00	0.60 ± 0.00	0.60 ± 0.00
		Flat Text	0.81 ± 0.00	0.70 ± 0.00	0.69 ± 0.00
	Transf.	Column	0.74 ± 0.00	0.65 ± 0.00	0.65 ± 0.00
		Flat Text	**0.86 ± 0.00**	**0.74 ± 0.00**	**0.75 ± 0.04**

Experimental Results – Without cross-table shift (1/2)

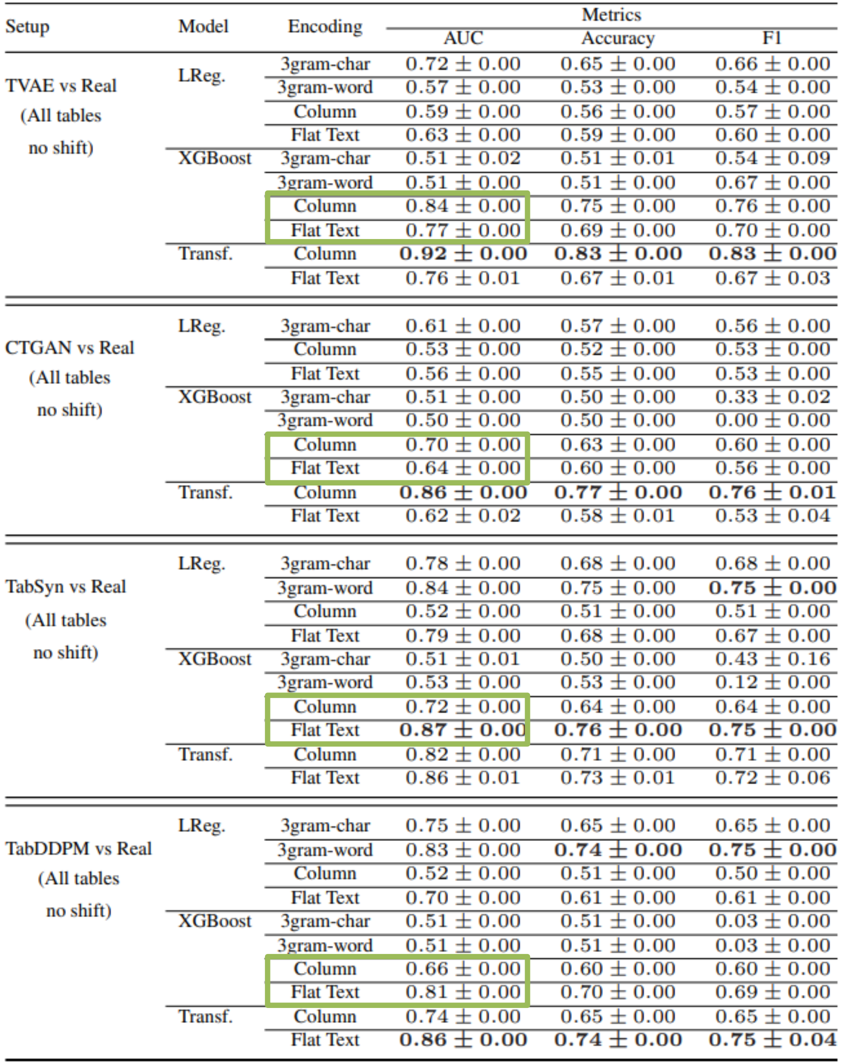
- Encoding matters, eg. XGBoost performance variation
- Poor performance with trigram encodings



Setup	Model	Encoding	Metrics		
			AUC	Accuracy	F1
TVAE vs Real (All tables no shift)	LReg.	3gram-char	0.72 ± 0.00	0.65 ± 0.00	0.66 ± 0.00
		3gram-word	0.57 ± 0.00	0.53 ± 0.00	0.54 ± 0.00
		Column	0.59 ± 0.00	0.56 ± 0.00	0.57 ± 0.00
		Flat Text	0.63 ± 0.00	0.59 ± 0.00	0.60 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.02	0.51 ± 0.01	0.54 ± 0.09
		3gram-word	0.51 ± 0.00	0.51 ± 0.00	0.67 ± 0.00
		Column	0.84 ± 0.00	0.75 ± 0.00	0.76 ± 0.00
		Flat Text	0.77 ± 0.00	0.69 ± 0.00	0.70 ± 0.00
	Transf.	Column	**0.92 ± 0.00**	**0.83 ± 0.00**	**0.83 ± 0.00**
		Flat Text	0.76 ± 0.01	0.67 ± 0.01	0.67 ± 0.03
CTGAN vs Real (All tables no shift)	LReg.	3gram-char	0.61 ± 0.00	0.57 ± 0.00	0.56 ± 0.00
		Column	0.53 ± 0.00	0.52 ± 0.00	0.53 ± 0.00
		Flat Text	0.56 ± 0.00	0.55 ± 0.00	0.53 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.00	0.50 ± 0.00	0.33 ± 0.02
		3gram-word	0.50 ± 0.00	0.50 ± 0.00	0.00 ± 0.00
		Column	0.70 ± 0.00	0.63 ± 0.00	0.60 ± 0.00
		Flat Text	0.64 ± 0.00	0.60 ± 0.00	0.56 ± 0.00
	Transf.	Column	**0.86 ± 0.00**	**0.77 ± 0.00**	**0.76 ± 0.01**
		Flat Text	0.62 ± 0.02	0.58 ± 0.01	0.53 ± 0.04
TabSyn vs Real (All tables no shift)	LReg.	3gram-char	0.78 ± 0.00	0.68 ± 0.00	0.68 ± 0.00
		3gram-word	0.84 ± 0.00	0.75 ± 0.00	**0.75 ± 0.00**
		Column	0.52 ± 0.00	0.51 ± 0.00	0.51 ± 0.00
		Flat Text	0.79 ± 0.00	0.68 ± 0.00	0.67 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.01	0.50 ± 0.00	0.43 ± 0.16
		3gram-word	0.53 ± 0.00	0.53 ± 0.00	0.12 ± 0.00
		Column	0.72 ± 0.00	0.64 ± 0.00	0.64 ± 0.00
		Flat Text	**0.87 ± 0.00**	**0.76 ± 0.00**	**0.75 ± 0.00**
	Transf.	Column	0.82 ± 0.00	0.71 ± 0.00	0.71 ± 0.00
		Flat Text	0.86 ± 0.01	0.73 ± 0.01	0.72 ± 0.06
TabDDPM vs Real (All tables no shift)	LReg.	3gram-char	0.75 ± 0.00	0.65 ± 0.00	0.65 ± 0.00
		3gram-word	0.83 ± 0.00	**0.74 ± 0.00**	**0.75 ± 0.00**
		Column	0.52 ± 0.00	0.51 ± 0.00	0.50 ± 0.00
		Flat Text	0.70 ± 0.00	0.61 ± 0.00	0.61 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.00	0.51 ± 0.00	0.03 ± 0.00
		3gram-word	0.51 ± 0.00	0.51 ± 0.00	0.03 ± 0.00
		Column	0.66 ± 0.00	0.60 ± 0.00	0.60 ± 0.00
		Flat Text	0.81 ± 0.00	0.70 ± 0.00	0.69 ± 0.00
	Transf.	Column	0.74 ± 0.00	0.65 ± 0.00	0.65 ± 0.00
		Flat Text	**0.86 ± 0.00**	**0.74 ± 0.00**	**0.75 ± 0.04**

Experimental Results – Without cross-table shift (1/2)

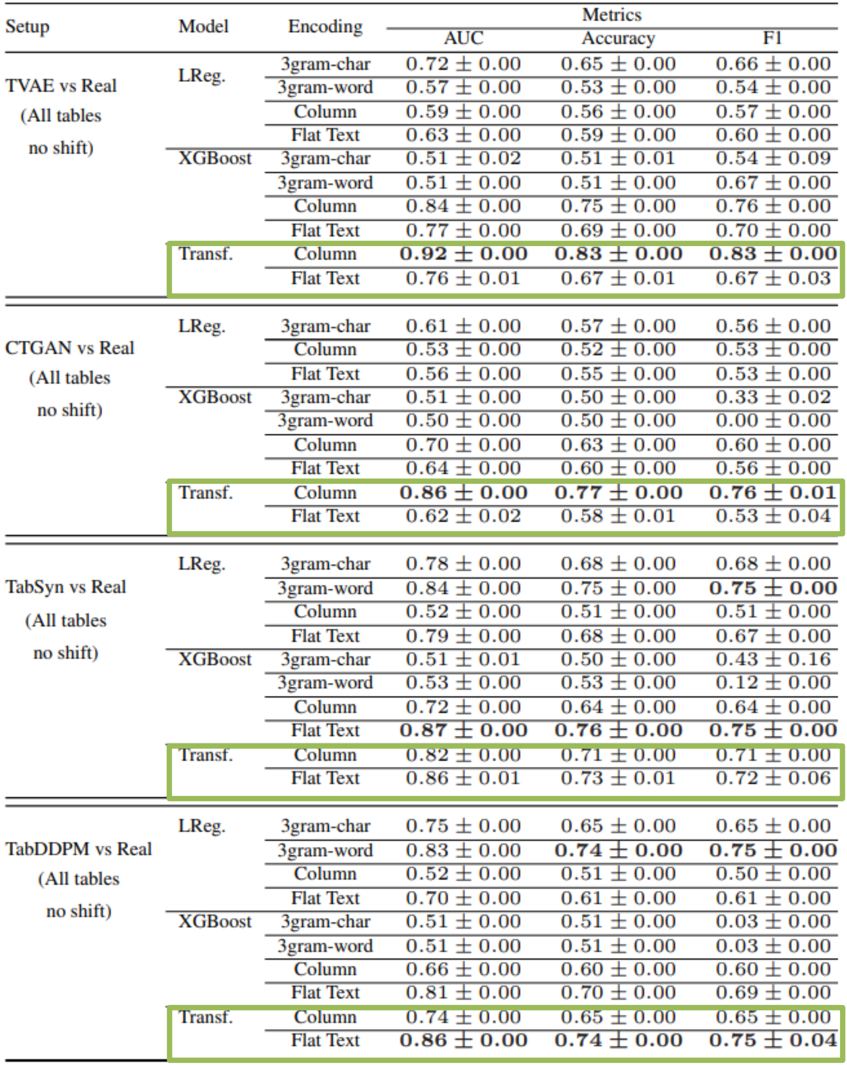
- Encoding matters, eg. XGBoost performance variation
- Poor performance with trigram encodings
- Good performance with other table-agnostic encodings



Setup	Model	Encoding	Metrics		
			AUC	Accuracy	F1
TVAE vs Real (All tables no shift)	LReg.	3gram-char	0.72 ± 0.00	0.65 ± 0.00	0.66 ± 0.00
		3gram-word	0.57 ± 0.00	0.53 ± 0.00	0.54 ± 0.00
		Column	0.59 ± 0.00	0.56 ± 0.00	0.57 ± 0.00
		Flat Text	0.63 ± 0.00	0.59 ± 0.00	0.60 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.02	0.51 ± 0.01	0.54 ± 0.09
		3gram-word	0.51 ± 0.00	0.51 ± 0.00	0.67 ± 0.00
		Column	0.84 ± 0.00	0.75 ± 0.00	0.76 ± 0.00
		Flat Text	0.77 ± 0.00	0.69 ± 0.00	0.70 ± 0.00
	Transf.	Column	**0.92 ± 0.00**	**0.83 ± 0.00**	**0.83 ± 0.00**
		Flat Text	0.76 ± 0.01	0.67 ± 0.01	0.67 ± 0.03
CTGAN vs Real (All tables no shift)	LReg.	3gram-char	0.61 ± 0.00	0.57 ± 0.00	0.56 ± 0.00
		Column	0.53 ± 0.00	0.52 ± 0.00	0.53 ± 0.00
		Flat Text	0.56 ± 0.00	0.55 ± 0.00	0.53 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.00	0.50 ± 0.00	0.33 ± 0.02
		3gram-word	0.50 ± 0.00	0.50 ± 0.00	0.00 ± 0.00
		Column	0.70 ± 0.00	0.63 ± 0.00	0.60 ± 0.00
		Flat Text	0.64 ± 0.00	0.60 ± 0.00	0.56 ± 0.00
	Transf.	Column	**0.86 ± 0.00**	**0.77 ± 0.00**	**0.76 ± 0.01**
		Flat Text	0.62 ± 0.02	0.58 ± 0.01	0.53 ± 0.04
TabSyn vs Real (All tables no shift)	LReg.	3gram-char	0.78 ± 0.00	0.68 ± 0.00	0.68 ± 0.00
		3gram-word	0.84 ± 0.00	0.75 ± 0.00	**0.75 ± 0.00**
		Column	0.52 ± 0.00	0.51 ± 0.00	0.51 ± 0.00
		Flat Text	0.79 ± 0.00	0.68 ± 0.00	0.67 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.01	0.50 ± 0.00	0.43 ± 0.16
		3gram-word	0.53 ± 0.00	0.53 ± 0.00	0.12 ± 0.00
		Column	0.72 ± 0.00	0.64 ± 0.00	0.64 ± 0.00
		Flat Text	**0.87 ± 0.00**	**0.76 ± 0.00**	**0.75 ± 0.00**
	Transf.	Column	0.82 ± 0.00	0.71 ± 0.00	0.71 ± 0.00
		Flat Text	0.86 ± 0.01	0.73 ± 0.01	0.72 ± 0.06
TabDDPM vs Real (All tables no shift)	LReg.	3gram-char	0.75 ± 0.00	0.65 ± 0.00	0.65 ± 0.00
		3gram-word	0.83 ± 0.00	**0.74 ± 0.00**	**0.75 ± 0.00**
		Column	0.52 ± 0.00	0.51 ± 0.00	0.50 ± 0.00
		Flat Text	0.70 ± 0.00	0.61 ± 0.00	0.61 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.00	0.51 ± 0.00	0.03 ± 0.00
		3gram-word	0.51 ± 0.00	0.51 ± 0.00	0.03 ± 0.00
		Column	0.66 ± 0.00	0.60 ± 0.00	0.60 ± 0.00
		Flat Text	0.81 ± 0.00	0.70 ± 0.00	0.69 ± 0.00
	Transf.	Column	0.74 ± 0.00	0.65 ± 0.00	0.65 ± 0.00
		Flat Text	**0.86 ± 0.00**	**0.74 ± 0.00**	**0.75 ± 0.04**

Experimental Results – Without cross-table shift (1/2)

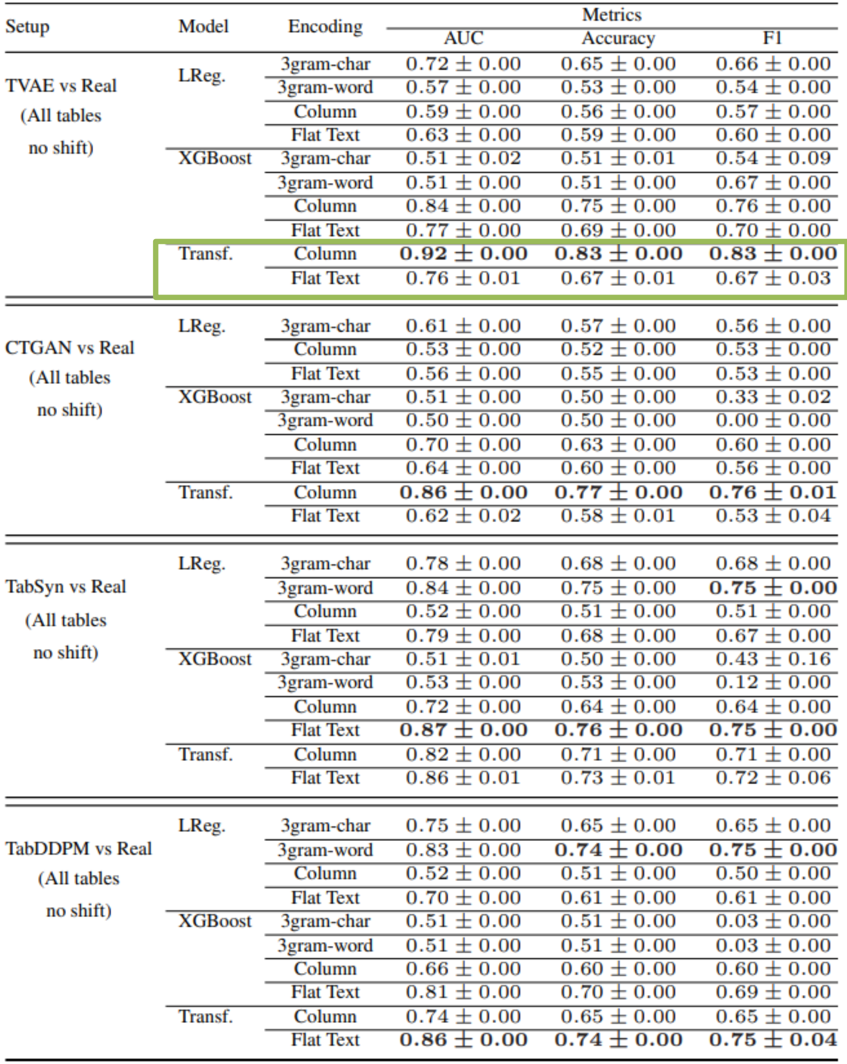
- Encoding matters, eg. XGBoost performance variation
- Poor performance with trigram encodings
- Good performance with other table-agnostic encodings
- Transformer-based detectors performance variation



Setup	Model	Encoding	Metrics		
			AUC	Accuracy	F1
TVAE vs Real (All tables no shift)	LReg.	3gram-char	0.72 ± 0.00	0.65 ± 0.00	0.66 ± 0.00
		3gram-word	0.57 ± 0.00	0.53 ± 0.00	0.54 ± 0.00
		Column	0.59 ± 0.00	0.56 ± 0.00	0.57 ± 0.00
		Flat Text	0.63 ± 0.00	0.59 ± 0.00	0.60 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.02	0.51 ± 0.01	0.54 ± 0.09
		3gram-word	0.51 ± 0.00	0.51 ± 0.00	0.67 ± 0.00
		Column	0.84 ± 0.00	0.75 ± 0.00	0.76 ± 0.00
		Flat Text	0.77 ± 0.00	0.69 ± 0.00	0.70 ± 0.00
	Transf.	Column	**0.92 ± 0.00**	**0.83 ± 0.00**	**0.83 ± 0.00**
		Flat Text	0.76 ± 0.01	0.67 ± 0.01	0.67 ± 0.03
CTGAN vs Real (All tables no shift)	LReg.	3gram-char	0.61 ± 0.00	0.57 ± 0.00	0.56 ± 0.00
		Column	0.53 ± 0.00	0.52 ± 0.00	0.53 ± 0.00
		Flat Text	0.56 ± 0.00	0.55 ± 0.00	0.53 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.00	0.50 ± 0.00	0.33 ± 0.02
		3gram-word	0.50 ± 0.00	0.50 ± 0.00	0.00 ± 0.00
		Column	0.70 ± 0.00	0.63 ± 0.00	0.60 ± 0.00
		Flat Text	0.64 ± 0.00	0.60 ± 0.00	0.56 ± 0.00
	Transf.	Column	**0.86 ± 0.00**	**0.77 ± 0.00**	**0.76 ± 0.01**
		Flat Text	0.62 ± 0.02	0.58 ± 0.01	0.53 ± 0.04
TabSyn vs Real (All tables no shift)	LReg.	3gram-char	0.78 ± 0.00	0.68 ± 0.00	0.68 ± 0.00
		3gram-word	0.84 ± 0.00	0.75 ± 0.00	**0.75 ± 0.00**
		Column	0.52 ± 0.00	0.51 ± 0.00	0.51 ± 0.00
		Flat Text	0.79 ± 0.00	0.68 ± 0.00	0.67 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.01	0.50 ± 0.00	0.43 ± 0.16
		3gram-word	0.53 ± 0.00	0.53 ± 0.00	0.12 ± 0.00
		Column	0.72 ± 0.00	0.64 ± 0.00	0.64 ± 0.00
		Flat Text	**0.87 ± 0.00**	**0.76 ± 0.00**	**0.75 ± 0.00**
	Transf.	Column	0.82 ± 0.00	0.71 ± 0.00	0.71 ± 0.00
		Flat Text	0.86 ± 0.01	0.73 ± 0.01	0.72 ± 0.06
TabDDPM vs Real (All tables no shift)	LReg.	3gram-char	0.75 ± 0.00	0.65 ± 0.00	0.65 ± 0.00
		3gram-word	0.83 ± 0.00	**0.74 ± 0.00**	**0.75 ± 0.00**
		Column	0.52 ± 0.00	0.51 ± 0.00	0.50 ± 0.00
		Flat Text	0.70 ± 0.00	0.61 ± 0.00	0.61 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.00	0.51 ± 0.00	0.03 ± 0.00
		3gram-word	0.51 ± 0.00	0.51 ± 0.00	0.03 ± 0.00
		Column	0.66 ± 0.00	0.60 ± 0.00	0.60 ± 0.00
		Flat Text	0.81 ± 0.00	0.70 ± 0.00	0.69 ± 0.00
	Transf.	Column	0.74 ± 0.00	0.65 ± 0.00	0.65 ± 0.00
		Flat Text	**0.86 ± 0.00**	**0.74 ± 0.00**	**0.75 ± 0.04**

Experimental Results – Without cross-table shift (2/2)

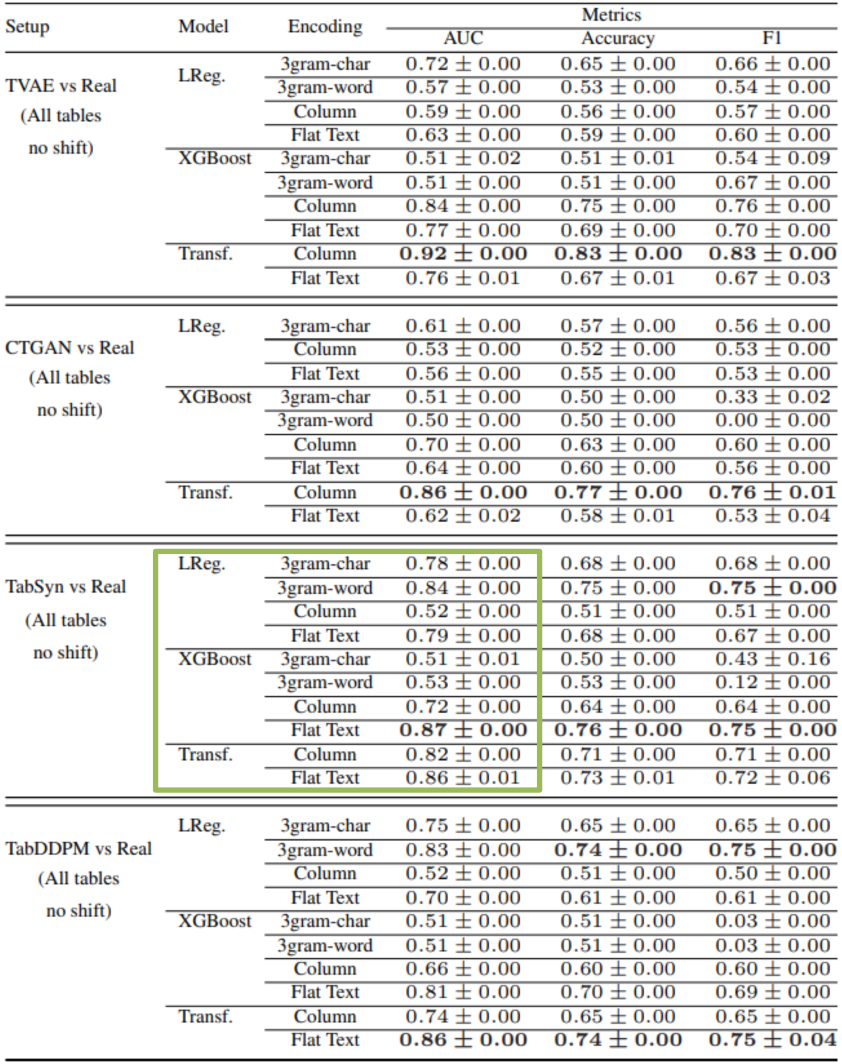
- TVAE is easy to detect



Setup	Model	Encoding	Metrics		
			AUC	Accuracy	F1
TVAE vs Real (All tables no shift)	LReg.	3gram-char	0.72 ± 0.00	0.65 ± 0.00	0.66 ± 0.00
		3gram-word	0.57 ± 0.00	0.53 ± 0.00	0.54 ± 0.00
		Column	0.59 ± 0.00	0.56 ± 0.00	0.57 ± 0.00
		Flat Text	0.63 ± 0.00	0.59 ± 0.00	0.60 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.02	0.51 ± 0.01	0.54 ± 0.09
		3gram-word	0.51 ± 0.00	0.51 ± 0.00	0.67 ± 0.00
		Column	0.84 ± 0.00	0.75 ± 0.00	0.76 ± 0.00
		Flat Text	0.77 ± 0.00	0.69 ± 0.00	0.70 ± 0.00
	Transf.	Column	**0.92 ± 0.00**	**0.83 ± 0.00**	**0.83 ± 0.00**
		Flat Text	0.76 ± 0.01	0.67 ± 0.01	0.67 ± 0.03
CTGAN vs Real (All tables no shift)	LReg.	3gram-char	0.61 ± 0.00	0.57 ± 0.00	0.56 ± 0.00
		Column	0.53 ± 0.00	0.52 ± 0.00	0.53 ± 0.00
		Flat Text	0.56 ± 0.00	0.55 ± 0.00	0.53 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.00	0.50 ± 0.00	0.33 ± 0.02
		3gram-word	0.50 ± 0.00	0.50 ± 0.00	0.00 ± 0.00
		Column	0.70 ± 0.00	0.63 ± 0.00	0.60 ± 0.00
		Flat Text	0.64 ± 0.00	0.60 ± 0.00	0.56 ± 0.00
	Transf.	Column	**0.86 ± 0.00**	**0.77 ± 0.00**	**0.76 ± 0.01**
		Flat Text	0.62 ± 0.02	0.58 ± 0.01	0.53 ± 0.04
TabSyn vs Real (All tables no shift)	LReg.	3gram-char	0.78 ± 0.00	0.68 ± 0.00	0.68 ± 0.00
		3gram-word	0.84 ± 0.00	0.75 ± 0.00	**0.75 ± 0.00**
		Column	0.52 ± 0.00	0.51 ± 0.00	0.51 ± 0.00
		Flat Text	0.79 ± 0.00	0.68 ± 0.00	0.67 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.01	0.50 ± 0.00	0.43 ± 0.16
		3gram-word	0.53 ± 0.00	0.53 ± 0.00	0.12 ± 0.00
		Column	0.72 ± 0.00	0.64 ± 0.00	0.64 ± 0.00
		Flat Text	**0.87 ± 0.00**	**0.76 ± 0.00**	**0.75 ± 0.00**
	Transf.	Column	0.82 ± 0.00	0.71 ± 0.00	0.71 ± 0.00
		Flat Text	0.86 ± 0.01	0.73 ± 0.01	0.72 ± 0.06
TabDDPM vs Real (All tables no shift)	LReg.	3gram-char	0.75 ± 0.00	0.65 ± 0.00	0.65 ± 0.00
		3gram-word	0.83 ± 0.00	**0.74 ± 0.00**	**0.75 ± 0.00**
		Column	0.52 ± 0.00	0.51 ± 0.00	0.50 ± 0.00
		Flat Text	0.70 ± 0.00	0.61 ± 0.00	0.61 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.00	0.51 ± 0.00	0.03 ± 0.00
		3gram-word	0.51 ± 0.00	0.51 ± 0.00	0.03 ± 0.00
		Column	0.66 ± 0.00	0.60 ± 0.00	0.60 ± 0.00
		Flat Text	0.81 ± 0.00	0.70 ± 0.00	0.69 ± 0.00
	Transf.	Column	0.74 ± 0.00	0.65 ± 0.00	0.65 ± 0.00
		Flat Text	**0.86 ± 0.00**	**0.74 ± 0.00**	**0.75 ± 0.04**

Experimental Results – Without cross-table shift (2/2)

- TVAE is easy to detect
- Good performance compared to [1] for a detection under the same table structure with XGBoost
- Eg. Average AUC on TabSyn in [1] = 0.63 vs 0.86 with our Text-based detector



Setup	Model	Encoding	Metrics		
			AUC	Accuracy	F1
TVAE vs Real (All tables no shift)	LReg.	3gram-char	0.72 ± 0.00	0.65 ± 0.00	0.66 ± 0.00
		3gram-word	0.57 ± 0.00	0.53 ± 0.00	0.54 ± 0.00
		Column	0.59 ± 0.00	0.56 ± 0.00	0.57 ± 0.00
		Flat Text	0.63 ± 0.00	0.59 ± 0.00	0.60 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.02	0.51 ± 0.01	0.54 ± 0.09
		3gram-word	0.51 ± 0.00	0.51 ± 0.00	0.67 ± 0.00
		Column	0.84 ± 0.00	0.75 ± 0.00	0.76 ± 0.00
		Flat Text	0.77 ± 0.00	0.69 ± 0.00	0.70 ± 0.00
	Transf.	Column	**0.92 ± 0.00**	**0.83 ± 0.00**	**0.83 ± 0.00**
		Flat Text	0.76 ± 0.01	0.67 ± 0.01	0.67 ± 0.03
CTGAN vs Real (All tables no shift)	LReg.	3gram-char	0.61 ± 0.00	0.57 ± 0.00	0.56 ± 0.00
		Column	0.53 ± 0.00	0.52 ± 0.00	0.53 ± 0.00
		Flat Text	0.56 ± 0.00	0.55 ± 0.00	0.53 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.00	0.50 ± 0.00	0.33 ± 0.02
		3gram-word	0.50 ± 0.00	0.50 ± 0.00	0.00 ± 0.00
		Column	0.70 ± 0.00	0.63 ± 0.00	0.60 ± 0.00
		Flat Text	0.64 ± 0.00	0.60 ± 0.00	0.56 ± 0.00
	Transf.	Column	**0.86 ± 0.00**	**0.77 ± 0.00**	**0.76 ± 0.01**
		Flat Text	0.62 ± 0.02	0.58 ± 0.01	0.53 ± 0.04
TabSyn vs Real (All tables no shift)	LReg.	3gram-char	0.78 ± 0.00	0.68 ± 0.00	0.68 ± 0.00
		3gram-word	0.84 ± 0.00	0.75 ± 0.00	**0.75 ± 0.00**
		Column	0.52 ± 0.00	0.51 ± 0.00	0.51 ± 0.00
		Flat Text	0.79 ± 0.00	0.68 ± 0.00	0.67 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.01	0.50 ± 0.00	0.43 ± 0.16
		3gram-word	0.53 ± 0.00	0.53 ± 0.00	0.12 ± 0.00
		Column	0.72 ± 0.00	0.64 ± 0.00	0.64 ± 0.00
		Flat Text	**0.87 ± 0.00**	**0.76 ± 0.00**	**0.75 ± 0.00**
	Transf.	Column	0.82 ± 0.00	0.71 ± 0.00	0.71 ± 0.00
		Flat Text	0.86 ± 0.01	0.73 ± 0.01	0.72 ± 0.06
TabDDPM vs Real (All tables no shift)	LReg.	3gram-char	0.75 ± 0.00	0.65 ± 0.00	0.65 ± 0.00
		3gram-word	0.83 ± 0.00	**0.74 ± 0.00**	**0.75 ± 0.00**
		Column	0.52 ± 0.00	0.51 ± 0.00	0.50 ± 0.00
		Flat Text	0.70 ± 0.00	0.61 ± 0.00	0.61 ± 0.00
	XGBoost	3gram-char	0.51 ± 0.00	0.51 ± 0.00	0.03 ± 0.00
		3gram-word	0.51 ± 0.00	0.51 ± 0.00	0.03 ± 0.00
		Column	0.66 ± 0.00	0.60 ± 0.00	0.60 ± 0.00
		Flat Text	0.81 ± 0.00	0.70 ± 0.00	0.69 ± 0.00
	Transf.	Column	0.74 ± 0.00	0.65 ± 0.00	0.65 ± 0.00
		Flat Text	**0.86 ± 0.00**	**0.74 ± 0.00**	**0.75 ± 0.04**

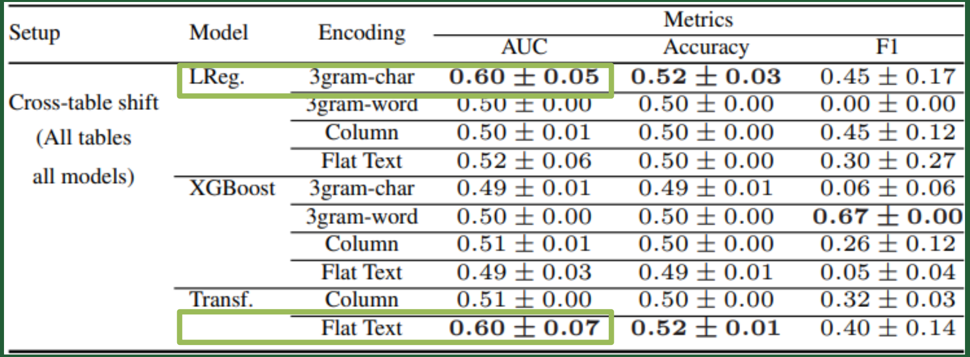
[1] G. Charbel N. Kindji, Lina Maria Rojas-Barahona, Elisa Fromont, Tanguy Urvoy. Under the Hood of Tabular Data Generation Models: Benchmarks with Extensive Tuning. 2024.

Experimental Results – With cross-table shift

Training and deploying the detectors on rows from distinct tables

Experimental Results – With cross-table shift (1/2)

- An extremely challenging problem
- However Text-based Transformer and Logistic Regression achieves an AUC of 0.60
- Potential of improvement, especially for the transformer-based approaches

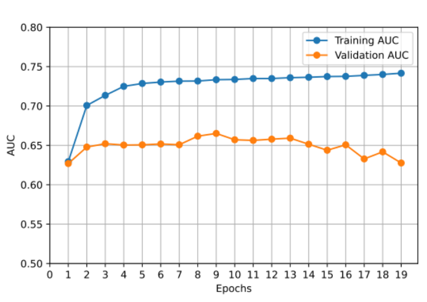


Setup	Model	Encoding	Metrics		
			AUC	Accuracy	F1
Cross-table shift (All tables all models)	LReg.	3gram-char	**0.60 ± 0.05**	**0.52 ± 0.03**	0.45 ± 0.17
		3gram-word	0.50 ± 0.00	0.50 ± 0.00	0.00 ± 0.00
		Column	0.50 ± 0.01	0.50 ± 0.00	0.45 ± 0.12
		Flat Text	0.52 ± 0.06	0.50 ± 0.00	0.30 ± 0.27
	XGBoost	3gram-char	0.49 ± 0.01	0.49 ± 0.01	0.06 ± 0.06
		3gram-word	0.50 ± 0.00	0.50 ± 0.00	**0.67 ± 0.00**
		Column	0.51 ± 0.01	0.50 ± 0.00	0.26 ± 0.12
		Flat Text	0.49 ± 0.03	0.49 ± 0.01	0.05 ± 0.04
	Transf.	Column	0.51 ± 0.00	0.50 ± 0.00	0.32 ± 0.03
		Flat Text	**0.60 ± 0.07**	**0.52 ± 0.01**	0.40 ± 0.14

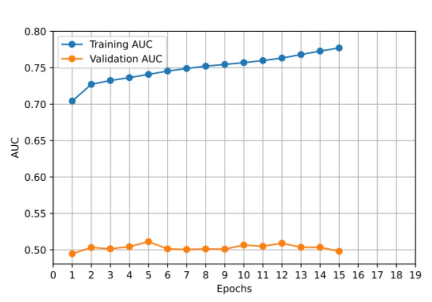
Experimental Results – With cross-table shift (2/2)

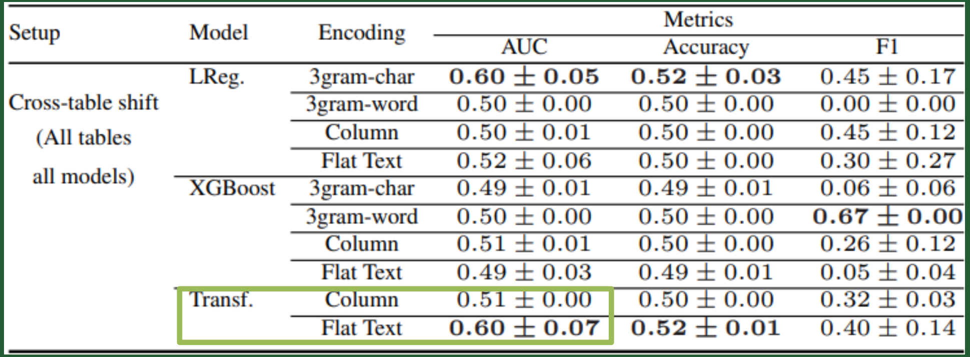
- Text-based transformer performing better than the column-based one during training as well
- Column-based transformer detector is overfiting

Text-based



Column-based





Setup	Model	Encoding	Metrics		
			AUC	Accuracy	F1
Cross-table shift (All tables all models)	LReg.	3gram-char	**0.60 ± 0.05**	**0.52 ± 0.03**	0.45 ± 0.17
		3gram-word	0.50 ± 0.00	0.50 ± 0.00	0.00 ± 0.00
		Column	0.50 ± 0.01	0.50 ± 0.00	0.45 ± 0.12
		Flat Text	0.52 ± 0.06	0.50 ± 0.00	0.30 ± 0.27
	XGBoost	3gram-char	0.49 ± 0.01	0.49 ± 0.01	0.06 ± 0.06
		3gram-word	0.50 ± 0.00	0.50 ± 0.00	**0.67 ± 0.00**
		Column	0.51 ± 0.01	0.50 ± 0.00	0.26 ± 0.12
		Flat Text	0.49 ± 0.03	0.49 ± 0.01	0.05 ± 0.04
	Transf.	Column	0.51 ± 0.00	0.50 ± 0.00	0.32 ± 0.03
		Flat Text	**0.60 ± 0.07**	**0.52 ± 0.01**	0.40 ± 0.14

Final Remarks

No cross-table shift → Good performance

Side result: good performance as compared to detection on the same table with ad hoc detector [1]

Cross-table shift → Very challenging problem

As expected, drop of performance but still AUC of 0.60 for the best detectors

Data encoding is key

Performance depends strongly on the data preprocessing scheme

Further investivation on transformers

Improvements to our results from the text-based transformer in recent work [2]

[1] G. Charbel N. Kindji, Lina Maria Rojas-Barahona, Elisa Fromont, Tanguy Urvoy. Under the Hood of Tabular Data Generation Models: Benchmarks with Extensive Tuning. 2024.

[2] G. Charbel N. Kindji, Elisa Fromont, Lina Maria Rojas-Barahona, Tanguy Urvoy. Datum-wise Transformer for Synthetic Tabular Data Detection in the Wild. 2025.