SnT

Interdisciplinary Centre for Security, Reliability and Trust

On the Robustness of Tabular Foundation Models

Dr. Karim Tit

ESSAI · European Cyber Week · November 2026

Interdisciplinary Centre for Security, Reliability and Trust



Dr. Karim Tit

Research associate

AI reliability and adversarial robustness
SerVal, SnT, University of Luxembourg

Our team members working on this topic:



Mohamed Djilani
PhD student



Thibault Simonetto
ML Engineer at POST



Florian Tambon
Postdoctoral Researcher



Salah Ghamizi
Research Scientist



Maxime Cordy
Associate Professor



Mike Papadakis
Associate Professor

Presenting today mostly the results from:

On the Robustness of Tabular Foundation Models: Test-Time Attacks and In-Context Defenses

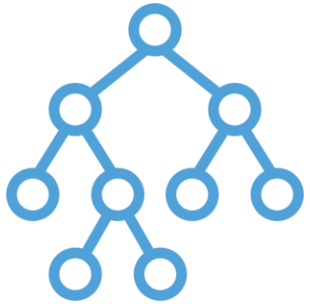
Accepted at IEEE SatML 2026

Agenda

1. Tabular foundation models and in-context learning
2. Evasion attacks and adversarial robustness
3. Adversarial In-Context Learning
4. Ongoing work and future directions

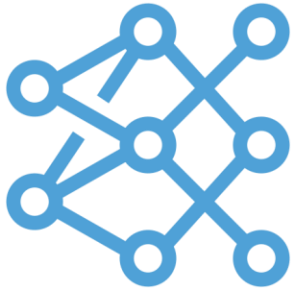
Background : Tabular Machine Learning History

A new paradigm shift ?



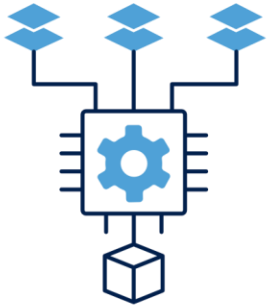
Tree-Based Models Domination (≈2000s–2021)

- Strong inductive bias
- *Interpretability*
- **GBDTs Dominates** Tab ML, particularly competitive DS



Neural Network Resurgence (≈2021–2024)

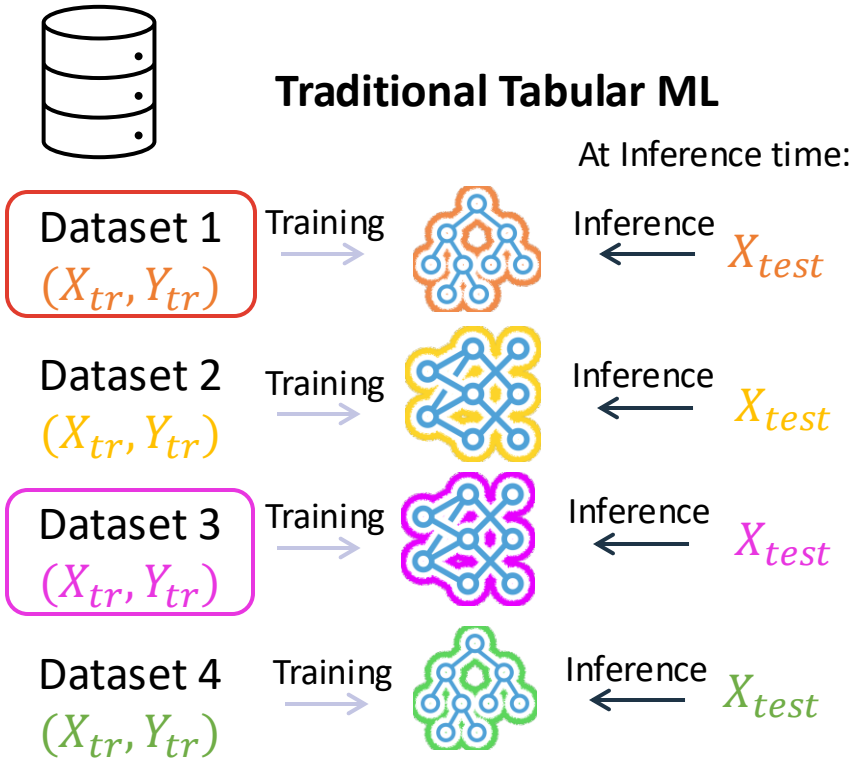
- Modalities Interactions
- **Strong performance** on very **large-scale** data
- Still dominated by GBDTs in small/medium data regimes

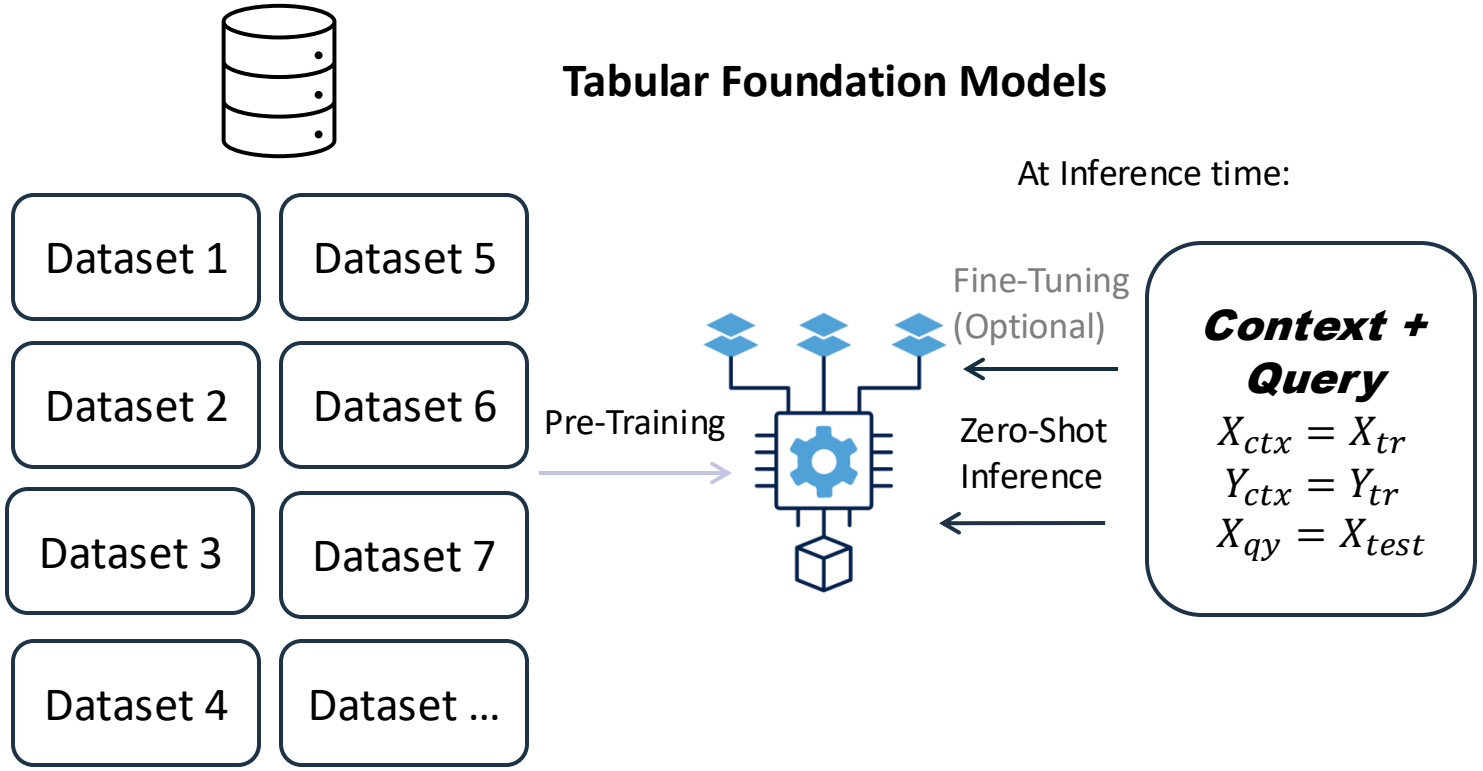


Tabular Foundation Model Era ? (≈2024–Present)

- Pre-trained on Large Synthetic tabular "dataset of datasets"
- **In-Context Learning** focus
- Dominant on **small/medium-scale** data

What are Tabular Foundation Models ?





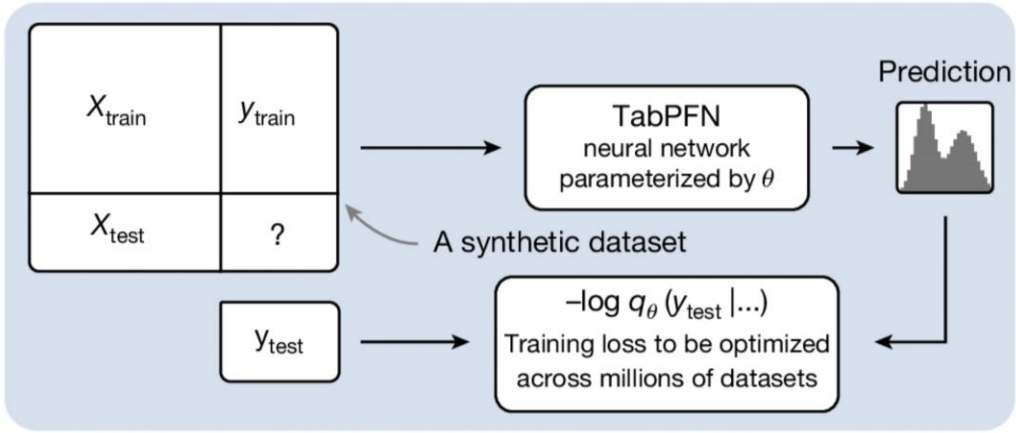
In Traditional ML the focus is on **Model Selection**:
→ Choosing the right model for a given dataset

With TFMs the focus moves to **Dataset and Architecture Selection**:
1. Choosing the right datasets for the **Pre-Training** phase
2. Choosing the right **global** architecture for **scalable** Inference

TabPFN: pretraining and prediction

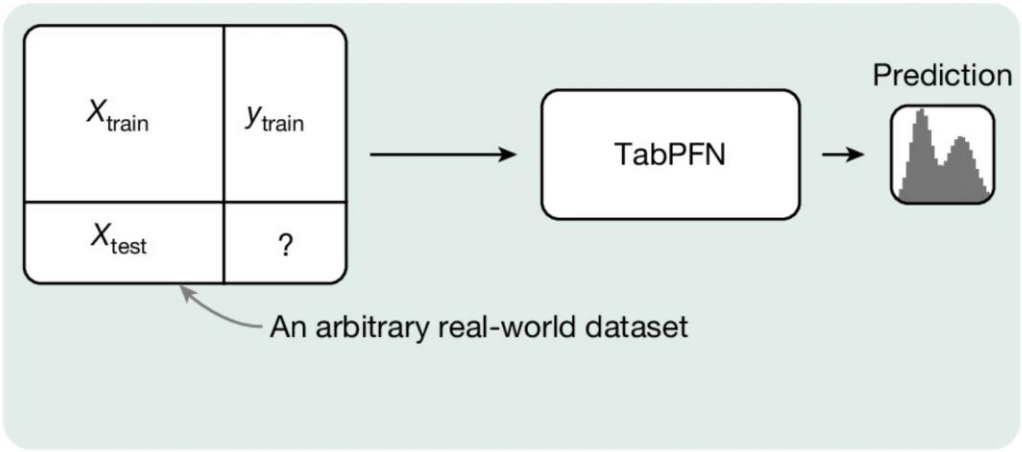
1) Pretraining

TabPFN is trained on synthetic data to take **entire datasets** as inputs and predict test labels in a **single forward pass**



2) Prediction

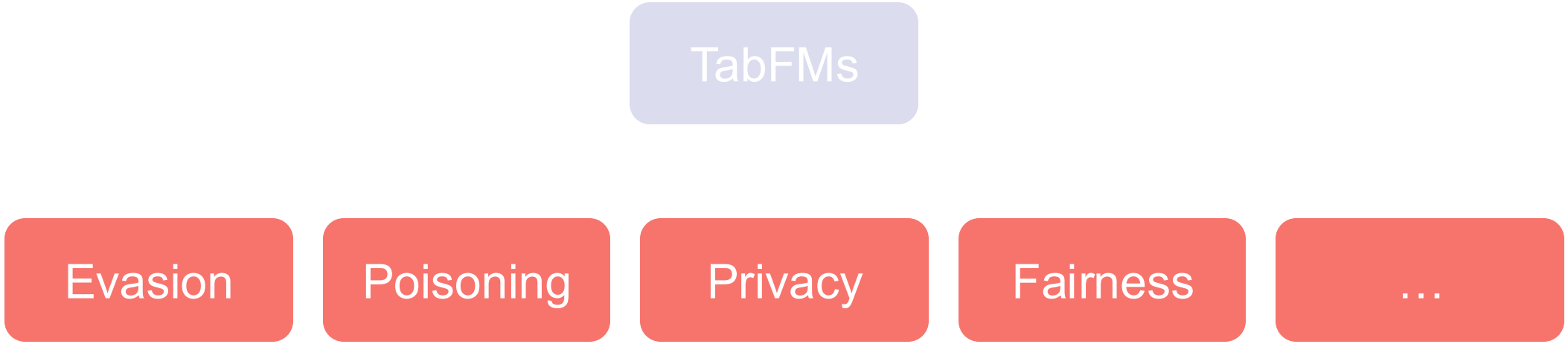
TabPFN can now be applied to arbitrary **unseen** real-world datasets



Hollmann, Noah, et al. "Accurate predictions on small data with a tabular foundation model." *Nature* 637.8045 (2025): 319-326.

Trustworthiness of TabFM

As Tabular Foundation Models adoption grows, so do concerns about its trustworthiness



Evasion : Adversarial manipulations of input features at inference time that subtly alter a data record (e.g., credit or medical data) to cause a model to output an incorrect prediction without changing the record's true underlying meaning.

Problem: Evasion Attacks on *Tabular* ML Models

Original

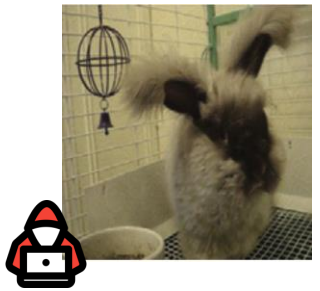


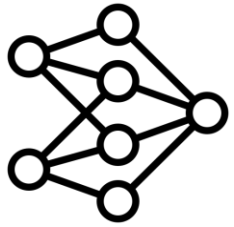
Angora: 91%

\+

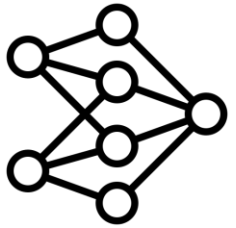
Bounded Perturbation →

Adversarial



→  → Strainer 77%

Age	Average Heart Rate	Resting Heart Rate	Calories	Active Minutes	Steps
72	135	78	210	45	5200

→ Constraint Validator →  →  →

Heart Condition
1

A Li, Yingwei, et al. "Learning transferable adversarial examples via ghost networks." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 07. 2020.

Problem: Evasion Attacks on Tabular ML Models

Manipulating Inputs to fool Tabular data ML Models

Age	Average Heart Rate	Resting Heart Rate	Calories	Active Minutes	Steps
72	135	78	210	45	5200



Age	Average Heart Rate	Resting Heart Rate	Calories	Active Minutes	Steps
72	110	78	210	50	5200



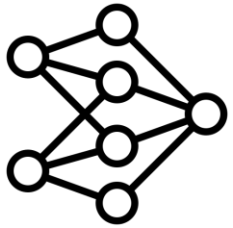


Constraint violated AHR≈ RHR×[1.5, 2.5]



Age	Average Heart Rate	Resting Heart Rate	Calories	Active Minutes	Steps
72	119	78	210	50	5200





Heart Condition
0

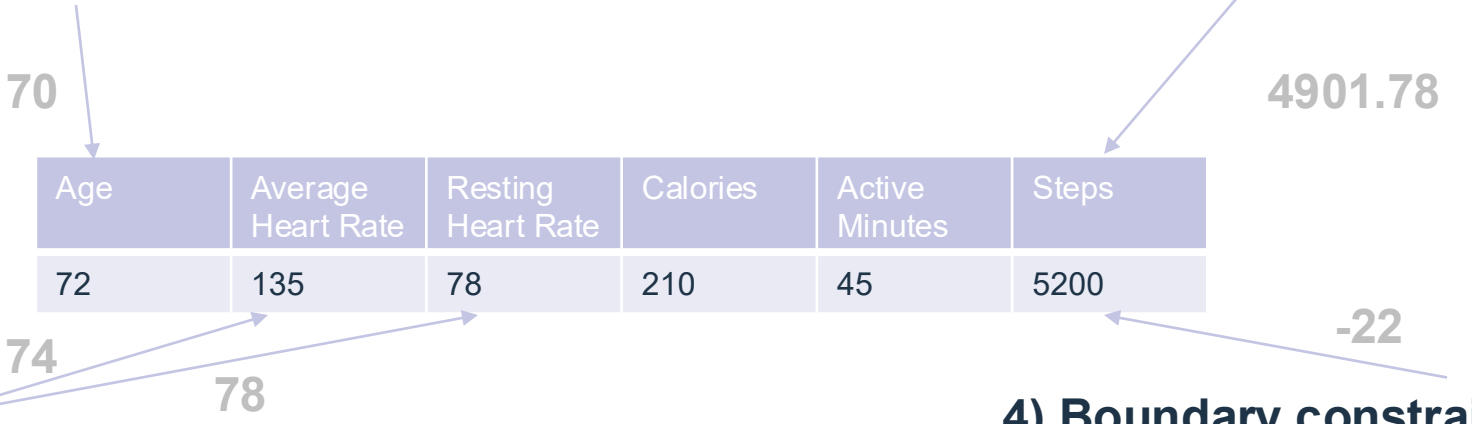
Flip the label
1 -> 0

Problem: Evasion Attacks on Tabular ML Models

Constraint Types:

1) Immutability constraints

2) Type constraints



3) Feature relationship constraints : AHR > RHR

4) Boundary constraints: Upper and lower bounds

Adversarial examples must remain valid and realistic within the domain of the tabular data.

Research Questions

1. How vulnerable are Tabular Foundation Models (TabFMs) such as TabPFNv2 and TabICL to evasion attacks ?

2. How can they be hardened against such threats?

How Vulnerable are TFMs ?
Setup

- *Adversary's Goal:*
  - Flip model predictions (e.g., Rejected → Approved)
  - Stay within perturbation budget L2 ($\varepsilon \leq 0.5$)

- *Adversary's Knowledge (Worst-Case):*
  - White-box access: knows model weights & context
  - Can compute gradients directly

- *Adversary's Capabilities:*
  - *Only modifies test inputs*
  - *Uses SoTA evasion attacks*
  - *Must respect domain constraints*

How Vulnerable are TFMs ?
Setup

Datasets

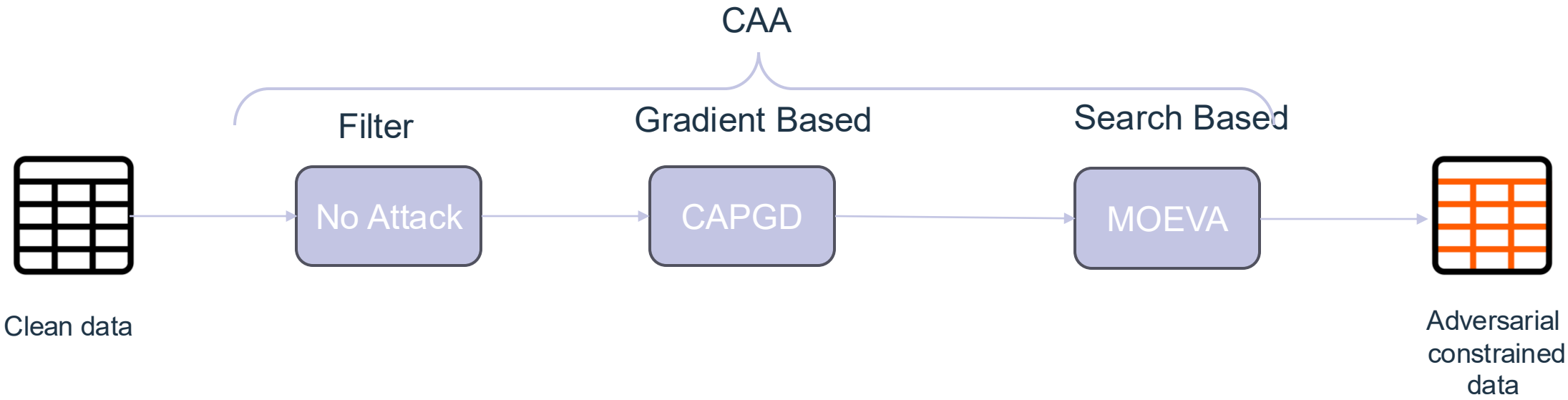
Dataset	Task	Properties Size	# Features	Balance
LCLD [31]	Credit Scoring	1 220 092	28	80/20
URL [32]	Phishing URL detection	11 430	63	50/50
WIDS [30]	ICU patient survival	91 713	186	91.4/8.6

Models

Family	Model
Foundational	TabPFNv2, TabICL
Transformer	TabTransformer
Transformer	TabNet
Regularization	RLN
Regularization	STG
Encoding	VIME
Tree-based	XGB
Tree-based	RF

How Vulnerable are TFMs ?
Setup

Adversarial evasion attack for tabular data: Constrained Adaptive Attack CAA



Metrics:

We consider the adversarial accuracy of the **positive class of interest, i.e. the recall**

$$\frac{1}{N}\sum_i [y_{1,true_i} = y_{pred_i}]$$

Simonetto, Thibault, Salah Ghamizi, and Maxime Cordy. "Constrained Adaptive Attack: Effective Adversarial Attack Against Deep Neural Networks for Tabular Data." Advances in Neural Information Processing Systems 37 (2024): 27817-27849.

How Vulnerable are TFMs ?
Experimental Results

	Dataset	LCLD			URL			WIDS		
	Model	Clean	CAPGD	CAA	Clean	CAPGD	CAA	Clean	CAPGD	CAA
Classical Models	RLN	68.1 (25.3)	0.2$_{\pm 0.1}$	0.0$_{\pm 0.1}$	94.4 (89.1)	11.5$_{\pm 0.3}$	10.6$_{\pm 0.4}$	77.5 (37.7)	62.1$_{\pm 0.4}$	60.9$_{\pm 0.4}$
	STG	66.0 (24.4)	56.1$_{\pm 0.3}$	52.6$_{\pm 0.3}$	93.3 (83.9)	71.5$_{\pm 0.2}$	56.9$_{\pm 0.3}$	77.6 (36.3)	64.2$_{\pm 0.6}$	63.6$_{\pm 0.7}$
	TabNet	67.5 (26.0)	7.3$_{\pm 0.4}$	0.3$_{\pm 0.2}$	93.4 (89.2)	17.9$_{\pm 0.9}$	9.1$_{\pm 0.4}$	79.7 (36.7)	7.9$_{\pm 0.3}$	5.5$_{\pm 0.5}$
	TabTr.	70.2 (25.2)	36.9$_{\pm 0.9}$	5.6$_{\pm 0.4}$	93.6 (88.0)	9.4$_{\pm 0.2}$	8.3$_{\pm 0.2}$	75.5 (38.5)	47.5$_{\pm 0.6}$	46.0$_{\pm 0.5}$
	VIME	67.1 (24.9)	3.6$_{\pm 0.4}$	2.8$_{\pm 0.4}$	92.5 (85.6)	54.4$_{\pm 0.2}$	47.1$_{\pm 1.3}$	72.3 (38.7)	50.7$_{\pm 0.3}$	50.3$_{\pm 0.3}$
	RF	64.6 (25.6)	N/A	27.5$_{\pm 0.6}$	95.9 (90.3)	N/A	48.3$_{\pm 1.6}$	49.3 (43.8)	N/A	11.0$_{\pm 0.4}$
	XGB	67.9 (26.1)	N/A	10.3$_{\pm 0.7}$	97.4 (93.8)	N/A	16.3$_{\pm 0.9}$	80.4 (40.5)	N/A	35.7$_{\pm 0.7}$
TabFMs	TabPFNv2	59.7 (25.4)	22.4 $_{\pm 0.1}$	9.0 $_{\pm 0.2}$	97.1 (92.8)	47.6 $_{\pm 0.2}$	25.2 $_{\pm 0.5}$	70.7 (43.0)	49.2 $_{\pm 0.4}$	30.8 $_{\pm 0.4}$
	TabICL	59.6 (25.5)	18.5 $_{\pm 0.4}$	10.2 $_{\pm 0.1}$	96.7 (92.8)	82.9 $_{\pm 0.2}$	20.2 $_{\pm 0.3}$	75.5 (42.0)	38.2 $_{\pm 0.3}$	30.5 $_{\pm 0.2}$

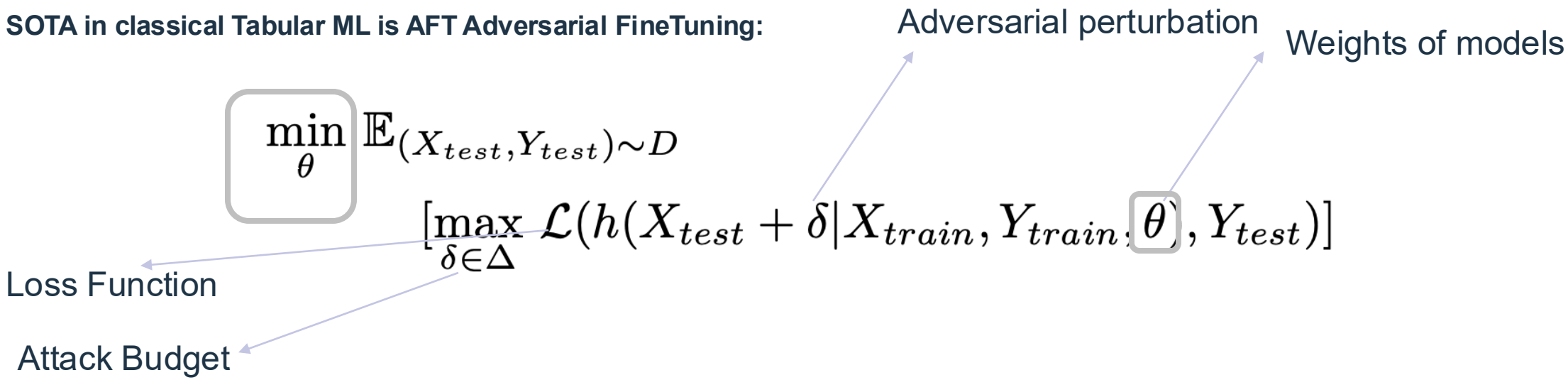
Recall against Clean data, CAPGD and CAA adversarial examples. For Clean data we also give Mathew Correlation Coefficient (MCC) in parenthesis. A **lower** robust performance means a more **effective** attack.

Answer to RQ1:

Our evidence shows that Tabular FMs are *at least* vulnerable to evasion attacks as their tabular non-FMs models counterparts.

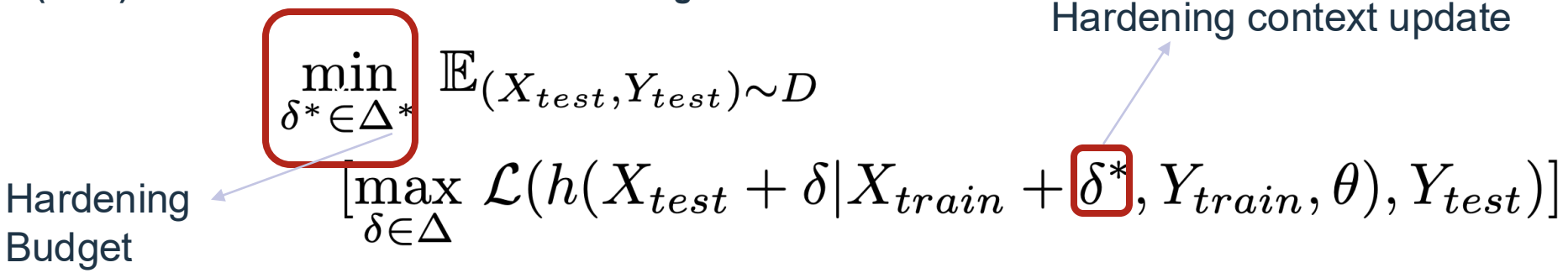
RQ2: Hardening of TabFMs: Problem Formulation

SOTA in classical Tabular ML is AFT Adversarial FineTuning:



The optimization is solved through stochastic gradient descent; usually on a mix of clean and attacked data

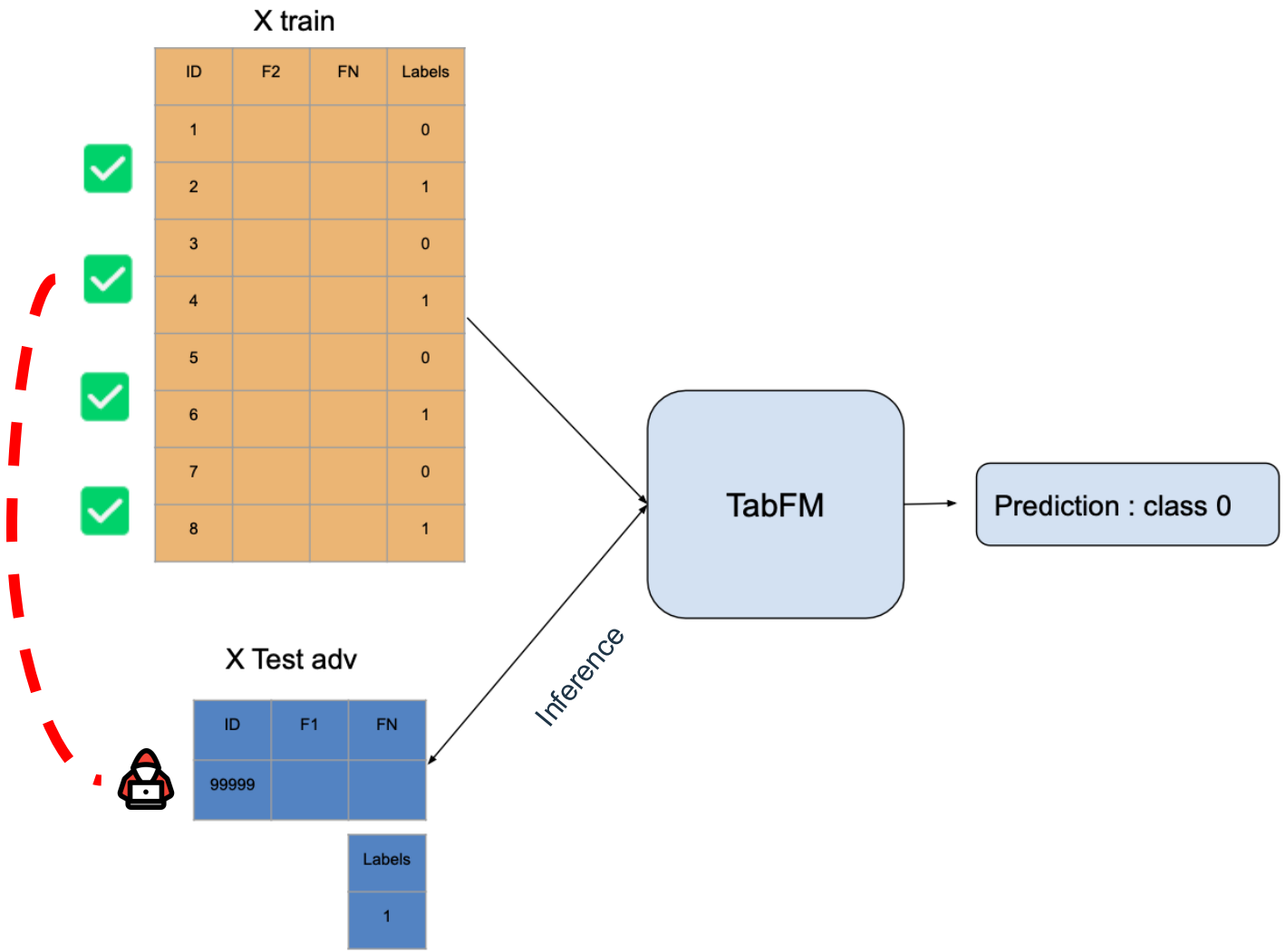
(Ours) AICL Adversarial In-Context-Learning:



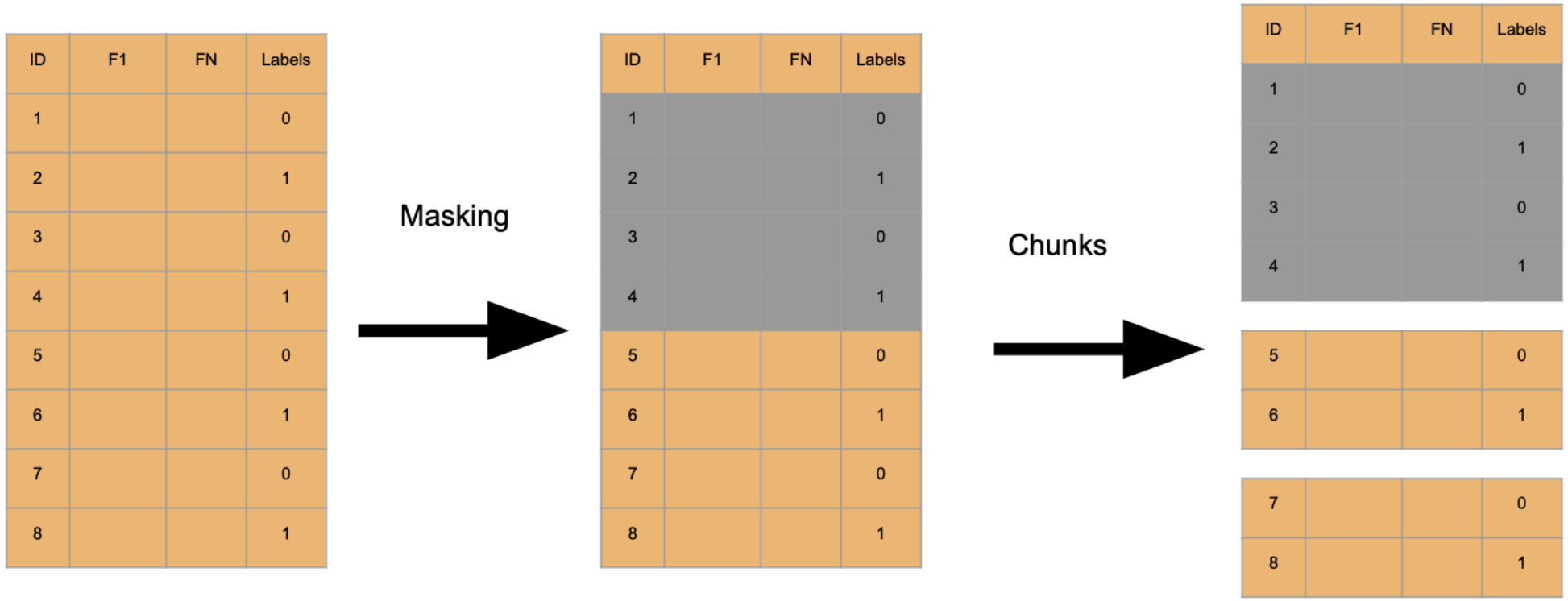
Hardening of TabFMs:
A Distribution Mismatch ?

TabFM hypothesis breakdown:

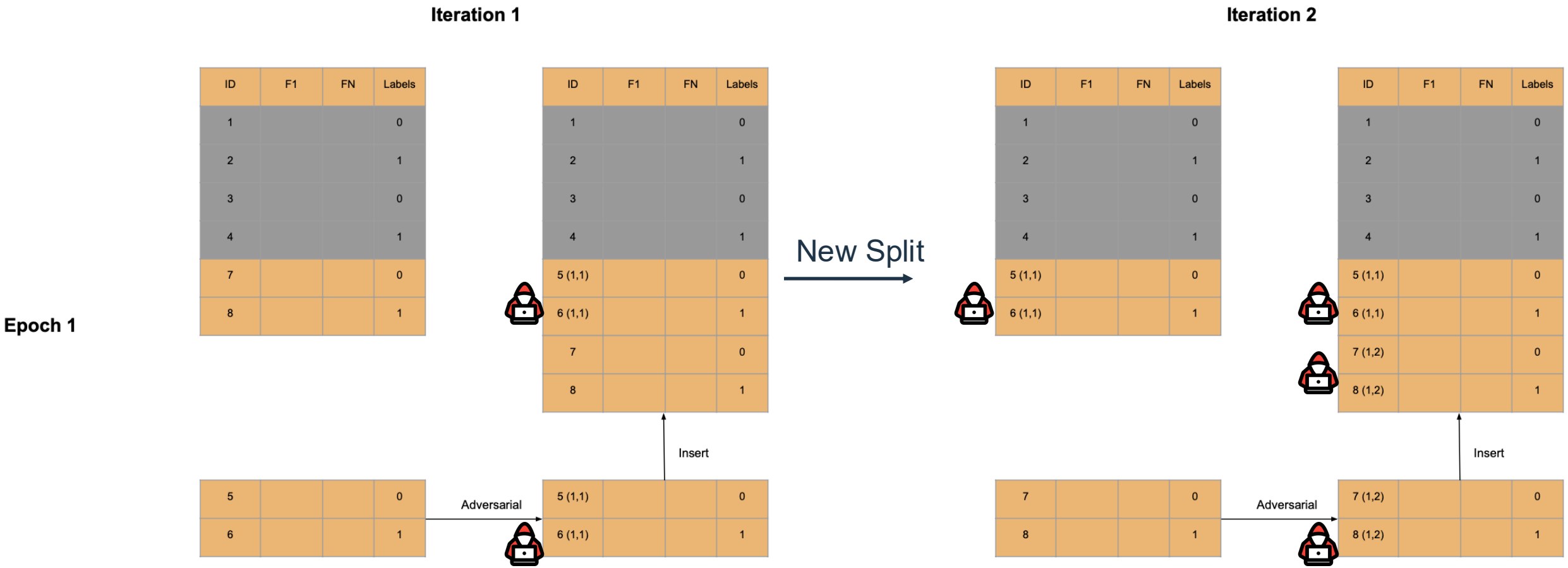
$$Dist(X_{ctx}) \neq Dist(X_{qy})$$



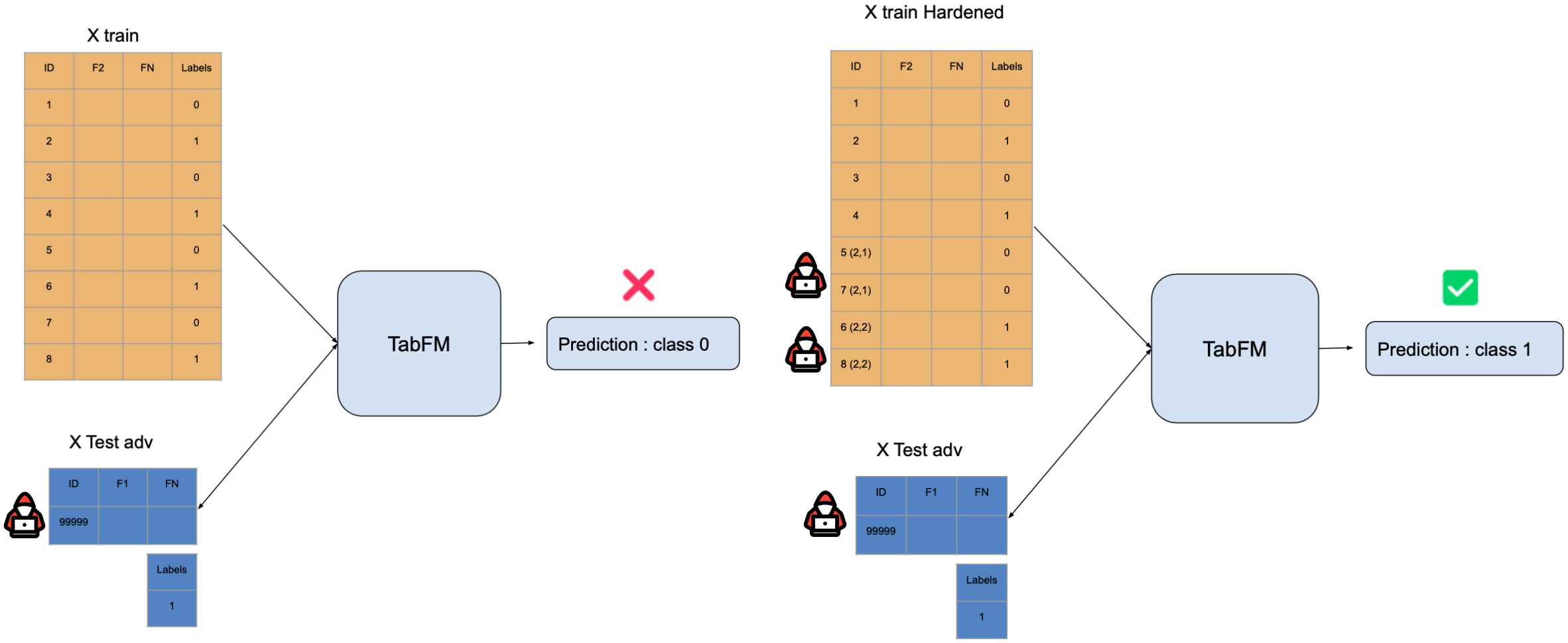
Hardening of TabFMs: the AICL Method (1/3)



Hardening of TabFMs: the AICL Method (2/3)



Hardening of TabFMs: The AICL Method (3/3)



Hardening of TabFMs: Results

Attacks used in hardening

Attacks used during evaluation

The Higher the Better

Dataset		LCLD			URL			WIDS		
Model	Training	Clean	CAPGD	CAA	Clean	CAPGD	CAA	Clean	CAPGD	CAA
TabPFN2	Original	59.7	22.4 ±0.1	9.0 ±0.2	97.1	47.6 ±0.2	25.2 ±0.5	70.7	49.2 ±0.4	30.8 ±0.4
	AICL CAA	60.4	49.3 ±0.3	41.2 ±0.0	96.2	96.1 ±0.0	69.4 ±0.2	70.7	65.7 ±0.3	51.5 ±0.2
	AICL CAPGD	55.8	42.0 ±0.5	14.7 ±0.2	95.6	94.7 ±0.1	25.9 ±0.2	70.7	63.7 ±0.2	42.5 ±0.4
	AFT CAA	62.7	20.6 ±0.4	12.0 ±0.1	97.2	51.1 ±0.2	31.2 ±0.3	73.3	38.2 ±0.3	31.1 ±0.5
	AFT CPGD	61.6	22.3 ±0.3	12.8 ±0.1	97.1	53.2 ±0.4	31.8 ±0.1	68.7	34.2 ±0.3	25.7 ±0.2
TabICL	Original	59.6	18.5 ±0.4	10.2 ±0.1	96.7	82.9 ±0.2	20.2 ±0.3	75.5	38.2 ±0.3	30.5 ±0.3
	AICL CAA	60.3	51.2 ±0.3	39.6 ±0.2	97.3	97.0 ±0.0	75.6 ±0.2	76.2	60.0 ±0.2	54.7 ±0.2
	AICL CAPGD	58.8	47.1 ±0.2	20.8 ±0.2	96.4	95.8 ±0.1	31.0 ±0.4	75.5	55.5 ±0.2	50.3 ±0.2
	AFT CAA	54.4	22.6 ±0.1	20.4 ±0.2	96.6	80.3 ±0.2	22.6 ±0.2	71.6	44.1 ±0.3	32.2 ±0.4
	AFT CPGD	59.2	18.4 ±0.2	4.3 ±0.1	96.7	84.8 ±0.3	16.8 ±0.1	54.6	35.2 ±0.4	14.8 ±0.3

Robust accuracy on Clean data and adversarial CAPGD/CAA attacks using AICL and AFT hardening methods

Answer to RQ2:

Robustness can be increased via the context only, by carefully replacing clean samples with their adversarial counterparts

Alternative: optimising context via gradients

With AICL, we replace context rows with adversarial training examples.

Alternatively one could directly optimises context features to reduce the loss on adversarial queries.

1. Generate adversarial queries against the current context.

2. Differentiate their loss with respect to selected context rows.

3. Project the updates into a bounded neighbourhood of the original context.

Future research directions

The context as an attack target

Context stealing

Can model outputs reveal private context examples?

Could gradients with respect to a candidate context support reconstruction?

Context poisoning

Can small changes to context examples steer predictions on future queries?

Explore context gradients under explicit attacker capabilities and constraints.

Advertising: We organize a workshop !

TrustTFM: Trustworthiness of Tabular Foundation Models

1st workshop at IEEE SaTML 2027 (3-5 May in Reykjavik, Iceland)

- How can we make tabular foundation models trustworthy in practice?
- Beyond predictive performance: **robustness**, **privacy,** and **interpretabitlity**

Our goal:
bringing together researchers and practitioners in tabular data and trustworthy AI.

- Discussing emerging threats, defenses and evaluation challenges.
- Interested in contributing? Join us and get i

Call for papers: deadline is 15th December.

Takeaways and discussion

- Tabular foundation models remain vulnerable to constrained evasion attacks.
- AICL improves robustness by adapting the context while keeping model weights fixed.
- GACT explores direct optimisation of context features against adversarial queries.
- The context also raises open questions about privacy and poisoning.

Our paper

My webpage