# Predicting and analyzing memorization within fine-tuned LLM

## Project Pitch

Jérémie Dentan[1], Davide Buscaldi[1, 2], Aymen Shabou[3], Sonia Vanier[1]

[1]*LIX (École Polytechnique, IP Paris, CNRS)*, [2]*LIPN (Sorbonne Paris Nord)*, [3]*Crédit Agricole SA*

### Acknowledgements

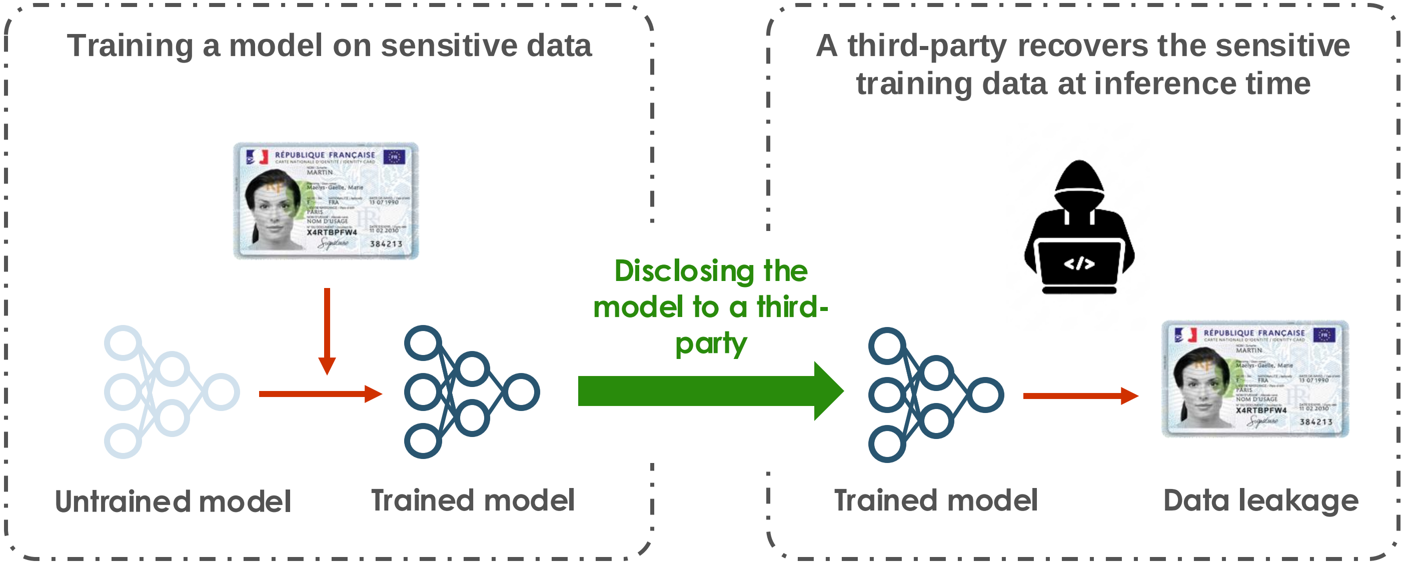
This work received financial support from Crédit Agricole SA through the research chair "Trustworthy and responsible AI" with École Polytechnique.

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011014843 made by GENCI.

Full paper: J Dentan, D Buscaldi, A Shabou, S Vanier. *Predicting and analyzing memorization within fine-tuned Large Language Models*. 2024. https://arxiv.org/abs/2409.18858

# Context: LLM memorize their training data

Large Language Models (LLM) memorize some training samples, which can be extracted at inference.




This can happen:

- **By mistake**, by anyone using the model
- **On purpose**, by an adversary willing to extract as much data as possible



Full paper: J Dentan, D Buscaldi, A Shabou, S Vanier. *Predicting and analyzing memorization within fine-tuned Large Language Models*. 2024. https://arxiv.org/abs/2409.18858

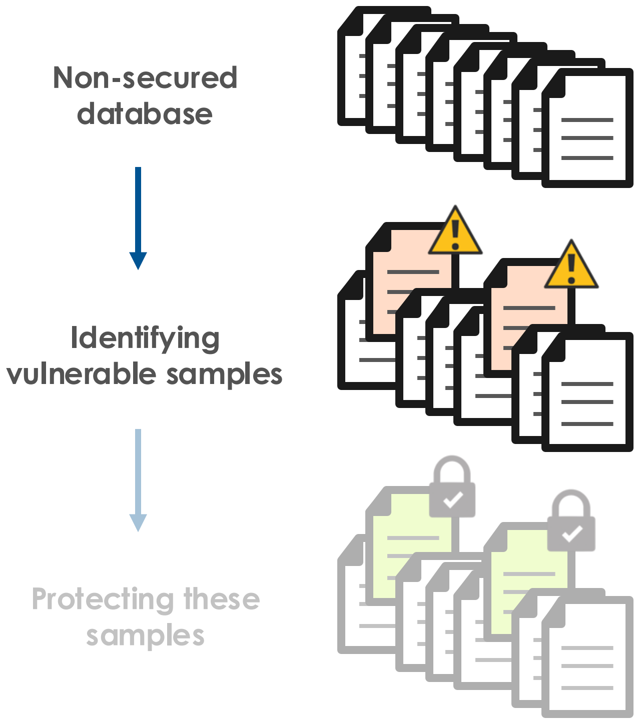
# Our goal: predicting memorization

We developed an auditing tool for practitioner to inspect models under development.

**Threat model:**

- A practitioner want to **audit a model under development** at minimal cost, before training.
- They perform some tests to identify vulnerable samples **before they are memorized.**
- The long-term objective is to protect these elements **at minimal cost.**

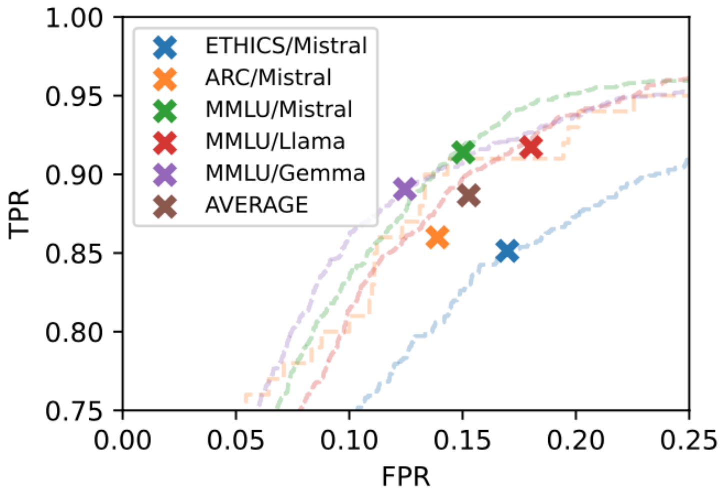
→ *In this paper, we only focus on the auditing tool.*
*The protection methods are left as future work.*



# Strong empirical results

We validated our approach in five different empirical settings, leading to strong results.

- We evaluated our approach on **Gemma 7B, Mistral 7B and Llama 2 7B** models fine-tuned for classification tasks (MMLU, ARC, ETHICS).
- We provide default hyperparameters for practitioners to **adapt to any classification task.**
- We obtain strong results: **FPR of 15.3% for a TPR of 88.7%.**




**CONCLUSION: Memorization can be predicted effectively from the early stages of training.**

Full paper: J Dentan, D Buscaldi, A Shabou, S Vanier. *Predicting and analyzing memorization within fine-tuned Large Language Models*. 2024. https://arxiv.org/abs/2409.18858