Trust and security in AI

ORAILIX team research projects @ LIX

Sonia VANIER and Jérémie DENTAN
École Polytechnique – LIX, ORAILIX team

Google Responsible AI Summit

Outline

I. Introduction
   A. Research topics
   B. Industrial partnerships
   C. Research impacts

II. Predicting LLM memorization
   A. Introduction
   B. Our approach
   C. Empirical results

Outline

I. Introduction
   A. Research topics
   B. Industrial partnerships
   C. Research impacts

II. Predicting LLM memorization
   A. Introduction
   B. Our approach
   C. Empirical results

Research topics: Hybrid approaches

Google Responsible AI Summit

We develop hybrid approaches between Artificial Intelligence (AI) and Operation Research (OR)

- **Real data** to strengthen OR solutions, scaling up, considering uncertainties
- **Reinforcement learning** for robustness and management of dynamic processes
- **Generative AI** to improve modelling, solver parametrization and generate better predictions

Modelling

- **Efficient modelling** of problems
- Provide **reliable, safe, explainable and optimal** solutions

Efficiency

- Integrate **structured knowledge**
- Integrate **business skills** into models

Frugality

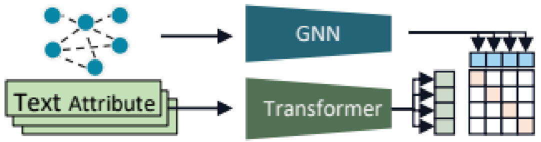
- **Reduce the size** of models and datasets
- **Minimize system resources** and vulnerabilities

Typical hybrid approaches

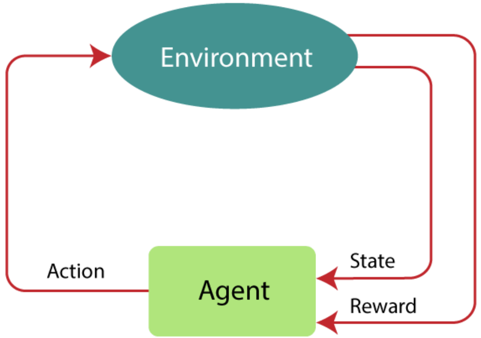
Google Responsible AI Summit

We work on hybrid approached for building more efficient models.

- Pipeline **Knowledge Graph (KG) + Language Model:**
  - Using KG as a **promising complement** to improve the quality of LLM with an external source of information.
  - Enhances information traceability, augmenting the model's explainability and **identify hallucinations.**



- Integrating **Reinforcement Learning and Operation Research:**
  - Particularly useful for **explainability** and training on limited (sensitive/private) data.
  - Ability to learn from data streams and **dynamic** processes.
  - **Uncertainty analysis** is crucial, for interpretability and security.



Research topics: Trustworthy and secured AI

Google Responsible AI Summit

École Polytechnique | Institut Polytechnique de Paris

We aim to develop responsible and trustworthy and secured AI systems

- Reduce the **environmental impact of AI**, its energy consumption and the size of models.
  → *Avoid training increasingly large models.*

- Respond to the **new security challenges** posed by AI-based systems and large-scale models,
  → *Vulnerabilities of companies using pre-trained models, such as hallucination.*

- Develop AI systems that help make decisions leading to **fair, equitable and ethical** outcomes.
  → *Consider the societal impact of AI.*

- **Mitigate biases** in the processing, while bringing **explicability, robustness and traceability** to the models.

Funded projects

Our team is involved in many projects funded by industrial partners

- **Crédit Agricole** (”Trustworthy and Responsible AI” chair)
- **SNCF** (“AI and Optimization for Mobility”)
- **Safran** (PhD funded via IRT-SystemX, Probabilistic estimation of health indicators for complex systems)
- **Renault** (PhD Cifre, Predictive maintenance of automotive production resources)
- **Orange** (PhD Cifre, Mathematical modeling and deployment optimization for Cloud architectures)

CRÉDIT AGRICOLE S.A. SNCF SAFRAN AEROSPACE · DEFENCE · SECURITY RENAULT orange

AI and Optimization for Mobility Chair

This research chair between École Polytechnique and SNCF was signed in September 2024.

Innovation to **optimize transport systems** and **sustainable mobility**

X-SNCF Research and Innovation **Partnership**

Trainings in sustainable mobility and transport for:
engineers, researchers, internships, phd theses...

Trustworthy and Responsible AI Chair

This research chair between École Polytechnique and Crédit Agricole was signed in November 2023.

- Long-term Research and Development on **trustworthy and responsible AI**

- **Research and education chair:** PhD students, post-docs, researchers, internship students

- **Project started:**
    - **PhD thesis** on trusted and responsible document processing models, and their applications to fraud detection
    - **PhD thesis** on data memorization within LLM and their robustness against model inversion attacks
    - **Post-doc** on differential privacy for protecting LLMs

CRÉDIT AGRICOLE S.A.

9

Impact of our research

We evaluate the economic, environmental and social impacts.

Cost reduction

More efficient management of financial resources.

Sustainability

Reducing greenhouse gas emissions and improving energy efficiency

Reinforced safety

Minimizing risks and improving safety.

An improved user experience

Ensure the performance of the products will benefit user experience.

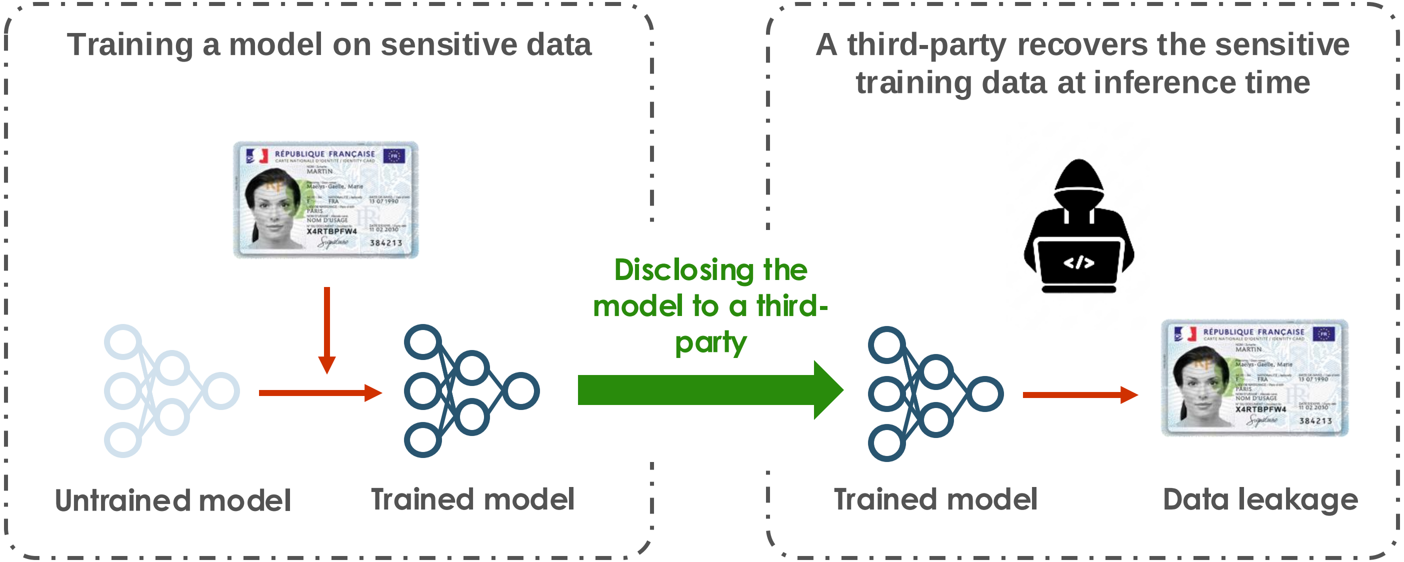
Outline

I. Introduction
   A. Research topics
   B. Industrial partnerships
   C. Research impacts

II. Predicting LLM memorization
   A. Introduction
   B. Our approach
   C. Empirical results

Scientific context: memorization in LLM

LLM memorize some training samples, which can be extracted at inference time.

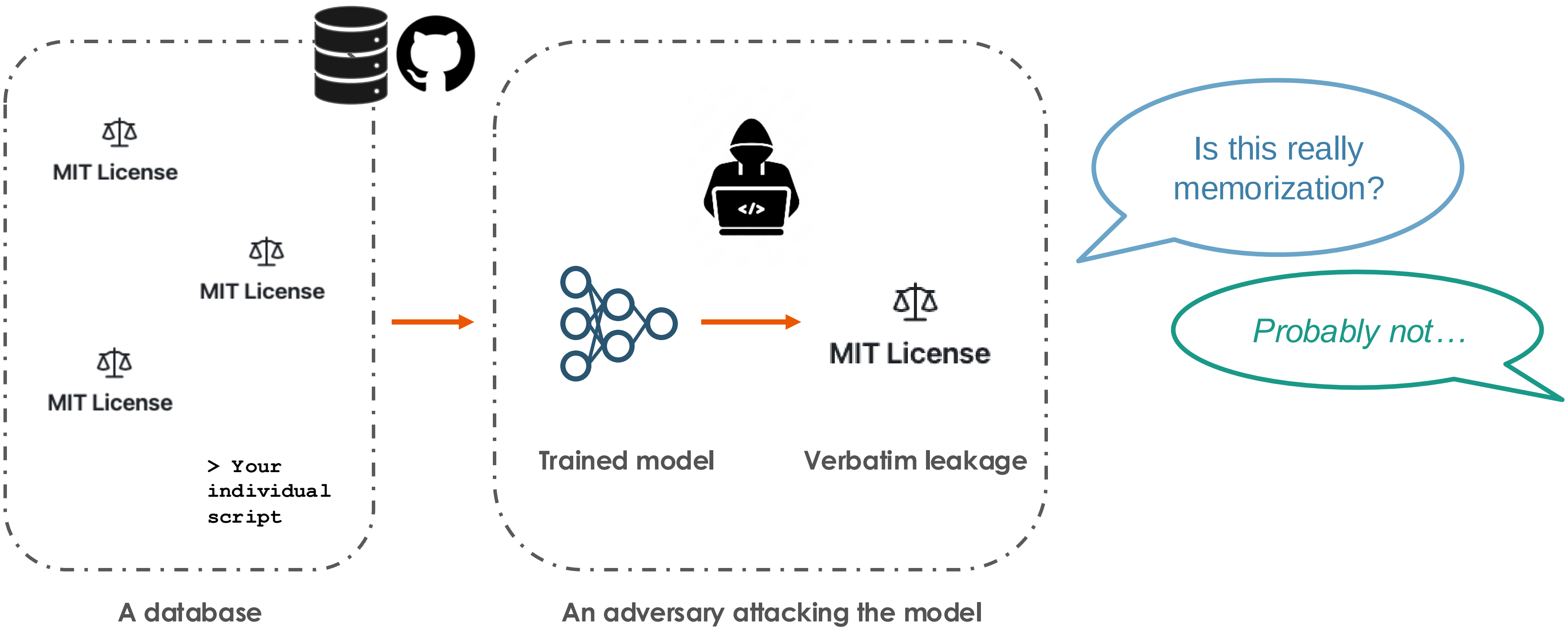


This can happen:

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

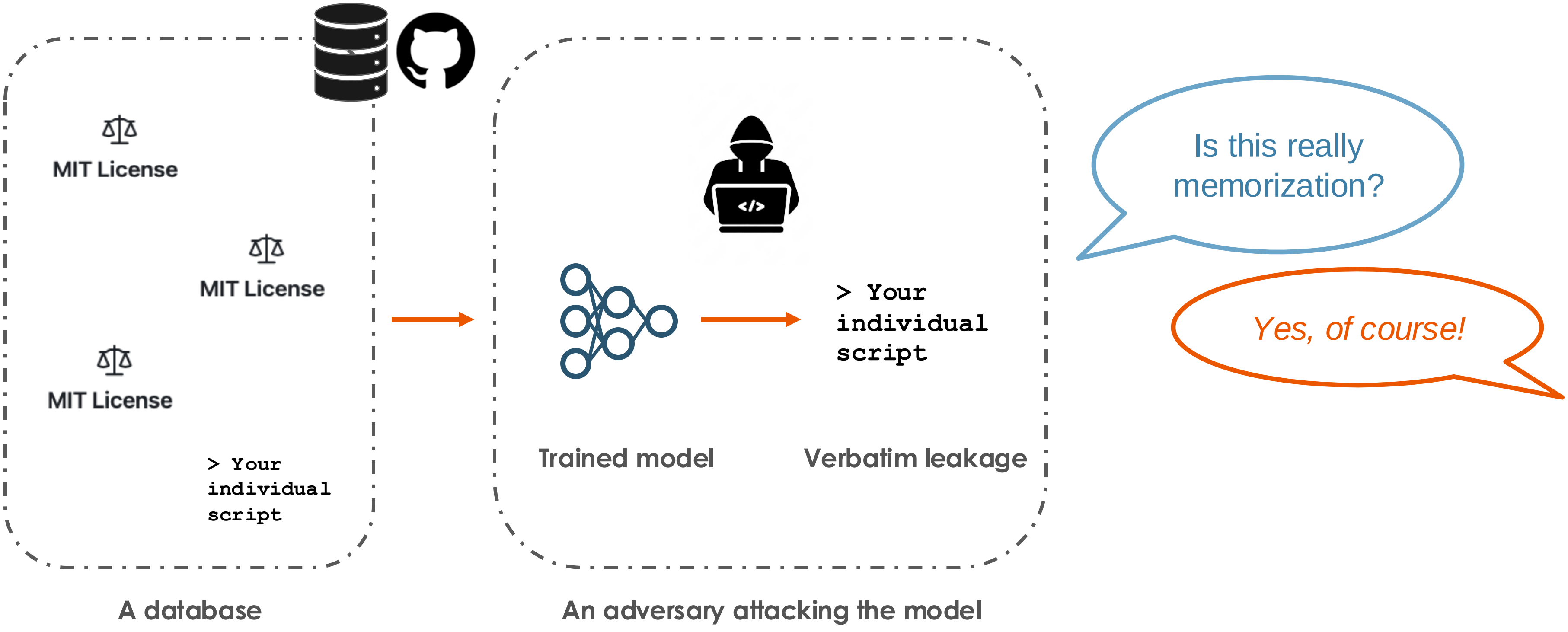
LLM memorization: an example

If a model outputs the MIT license verbatim, should I consider it as memorization?



LLM memorization: an example

If a model outputs the MIT license verbatim, should I consider it as memorization?



Defining memorization in LLM

Google Responsible AI Summit

L'École Polytechnique | Institut Polytechnique de Paris

Memorization is a complex concept, for which there exist many definitions.

Extractability

Is it possible to extract this sample from the model, using an adversarial attack?

Differential Privacy

A theoretical measure to bound the information the adversary can obtain.

Membership Inference

Can an adversary guess if a sample was part of the training set of the model?

Counterfactual memorization

What it the individual impact of each sample on the weights of the model?

➔ **As a practitioner, how do I audit my model to know if it has "memorized" training data?**

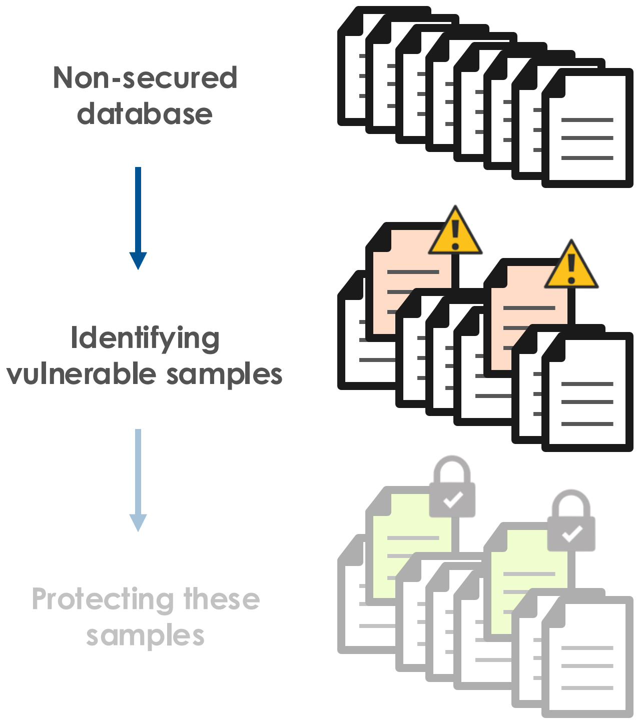
Predicting memorization

We aim to develop 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.*

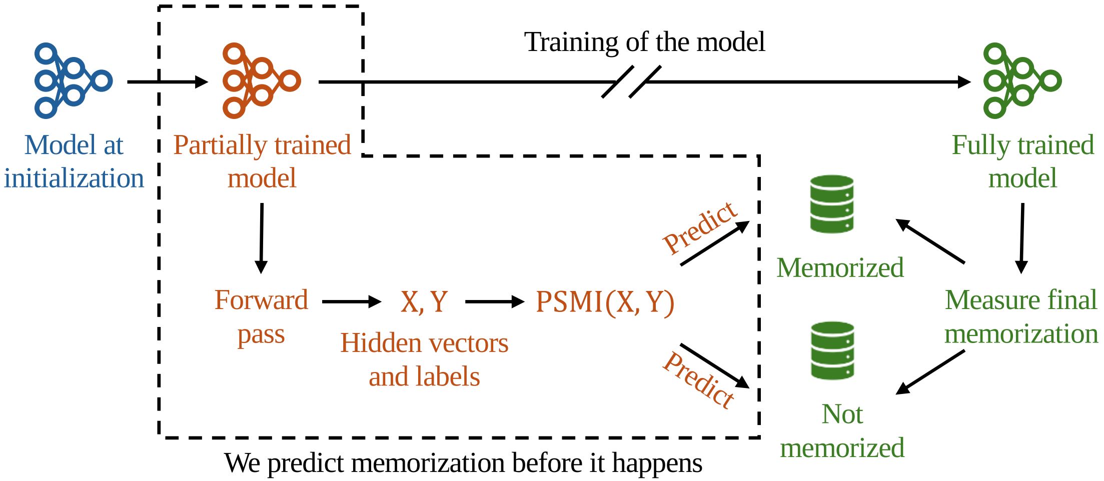


[1] J Dentan, D Buscaldi, A Shabou, S Vanier. *Predicting and analysing memorization within fine-tuned Large Language Models*. Arxiv preprint. 2024.

Overview of our pipeline

Google Responsible AI Summit

We interrupt training at the initial stages to predict memorization before it happens.



PSMI = Pointwise Sliced Mutual Information [2]
= how surprising label Y is when we observe hidden vector X

Key points:

- We predict memorization **before it happens**
- Easy to compute, with a **realistic budget.**
- Supported by theoretical results, and **easily adaptable** to any classification problem.

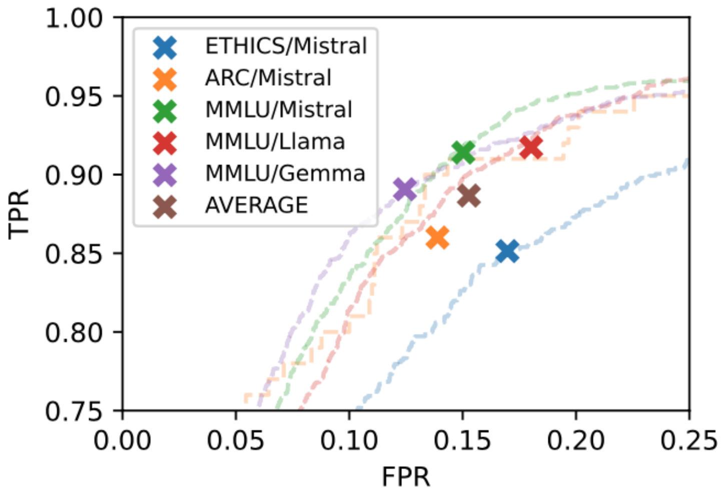
[1] J Dentan, D Buscaldi, A Shabou, S Vanier. *Predicting and analysing memorization within fine-tuned Large Language Models*. Arxiv preprint. 2024.
[2] S Wongso, R Ghosh, M Motani. *Pointwise Sliced Mutual Information for Neural Network Explainability*. IEEE ISIT. 2023.

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%.**

→ *In the paper: other analysis: impact of the hyperparameters, comparison with existing baselines, etc.*



[1] J Dentan, D Buscaldi, A Shabou, S Vanier. *Predicting and analysing memorization within fine-tuned Large Language Models*. Arxiv preprint. 2024.

Thank you

Acknowledgements.

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

This work received financial support from SNCF through the research chair ”AI and Optimization for Mobility" with École Polytechnique.

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