Comprendre et mesurer les risques de confidentialité des modèles de langage

Sonia VANIER et Jérémie DENTAN – École Polytechnique

Travaux réalisés au sein de la Chaire « IA de confiance et responsable » Polytechnique & Crédit Agricole
Forum Industriel de l'IA 2025 de l'AFIA

Chaire « IA de Confiance et Responsable »

- Objectifs de la chaire
- Thématiques de recherche
- Enjeux
- L’équipe Orailix
- Partenariats
- MScT "Trustworthy and Responsible AI"

Chaire *IA de Confiance et Responsable*

Une chaire de recherche a été signée fin 2023 entre Polytechnique et le Crédit Agricole

- Projets de R&D long-terme sur **l'IA de Confiance et Responsable**
- **Chaire de recherche et d'éducation:** doctorants, post-docs, chercheurs, stagiaires
- **Projets en cours:**
  - **Doctorat:** modèles de traitement fiables et responsable avec application à la détection de fraude.
  - **Doctorat:** mémorisation des données dans les LLM
  - **Doctorat:** systèmes de recommandation fiable, multimodaux et explicables.
  - **Post-doc:** confidentialité différentielle appliquée aux LLM
  - Stages à venir…

Axes de recherche

Les projets de recherche de la chaire sont structurés autour de différents axes.

- Répondre aux **nouveaux enjeux de sécurité** posés par les systèmes d'IA et les modèles à grande échelle
  → *Fiabilité des modèles, confidentialité, robustesses aux cyber-attaques*

- **Mitiger les biais** des modèles, viser des modèles **explicables, robustes et traçables.**
  → *Conformité vis-à-vis de la réglementation des standards de qualité du groupe*

- Développer des systèmes d'IA aidant à prendre des décisions **justes, équitables et éthiques.**
  → *Application à la fraude bancaire, impact sociétal de l'IA*

- Réduire **l'impact environnemental** de l'IA et les consommations énergétiques associées
  → *Algorithmes performants et frugaux en calculs*

Approches hybrides

Nous développons des approches hybrides entre l'IA et la Recherche Opérationnelle

- **Données réelles** pour renforcer des systèmes de RO, passer à l'échelle et gérer les incertitudes.
- **Du reinforcement learning** pour la robustesse et la gestion de processus dynamiques
- **Des IA génératives** pour améliorer la modélisation et l'amélioration des prédictions générées.

Modélisation efficace

- **Modélisation efficace** des problèmes
- Trouver des solutions **fiables, sûres, optimales et explicables.**

Performance

- Intégration de **connaissances structurées.**
- Intégration de **connaissances métier.**

Frugalité

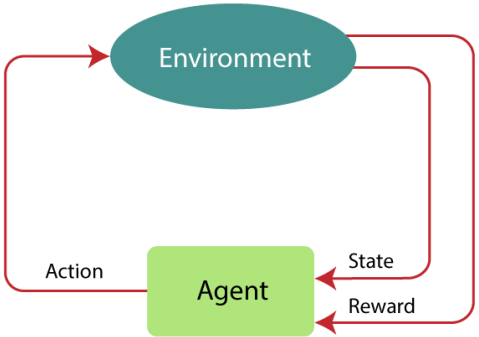
- **Réduction de la taille** des modèles et datasets.
- **Algorithmes efficaces** réduisant les temps de calcul.

Approches typiques

Nous développons des systèmes plus efficaces via des approches hybrides

- Pipeline **Graphe de connaissance + Modèle de langage :**
  - Utiliser un graphe de connaissance comme complément pour améliorer la qualité d'un LLM en tant que **source d'information extérieure.**
  - Améliorer la **traçabilité de l'information**, augmenter l'explicabilité et identifier les hallucinations.

- Combiner **Reinforcement Learning et Recherche Opérationnelle:**
  - Utile pour **l'explicabilité** et l'entraînement sur des données limitées.
  - Apprendre à partir de flux de données sur des **systèmes dynamiques.**
  - **Analyse des incertitudes**, notamment pour l'interprétabilité et la sécurité



Impacts à long terme

Nous évaluons les impacts économiques, environnementaux et sociaux de notre recherche.

Réduction des coûts

Gestion plus efficace des ressources financières

Développement durable

Réduction des émissions de CO2 et de la consommation énergétique

Sécurité renforcée

Minimiser les risques et améliorer la sécurité

Expérience uitilisateur

S'assurer que les produits bénéficieront aux utilisateurs.

Équipe Polytechnique: Orailix

Côté Polytechnique, la chaire est portée par l'équipe ORAILIX: **O**perations **R**esearch, **AI** @ **LIX**

ORAILIX

Recrutements: ouverture de deux postes de maître de conférence dans l'équipe!

Un environnement riche en collaborations

L'équipe ORAILIX est impliquée dans des projets financés par de multiples partenaires.

Crédit Agricole (chaire « IA de Confiance et Responsable »)

SNCF (chaire « IA et Optimisation pour les Mobilités »)

ORAILIX

Safran (thèse financée via IRT-SystemX sur l'estimation de l'état de santé de systèmes complexes)

Orange (thèse Cifre sur la modélisation et l'optimisation du déploiement sur le Cloud)

Renault (thèse Cifre sur la maintenance prédictive des ressources de production automobiles

MScT « Trustworthy and Responsible AI»

Rentrée 2025: Master international de haut niveau, spécifique à l'École Polytechnique

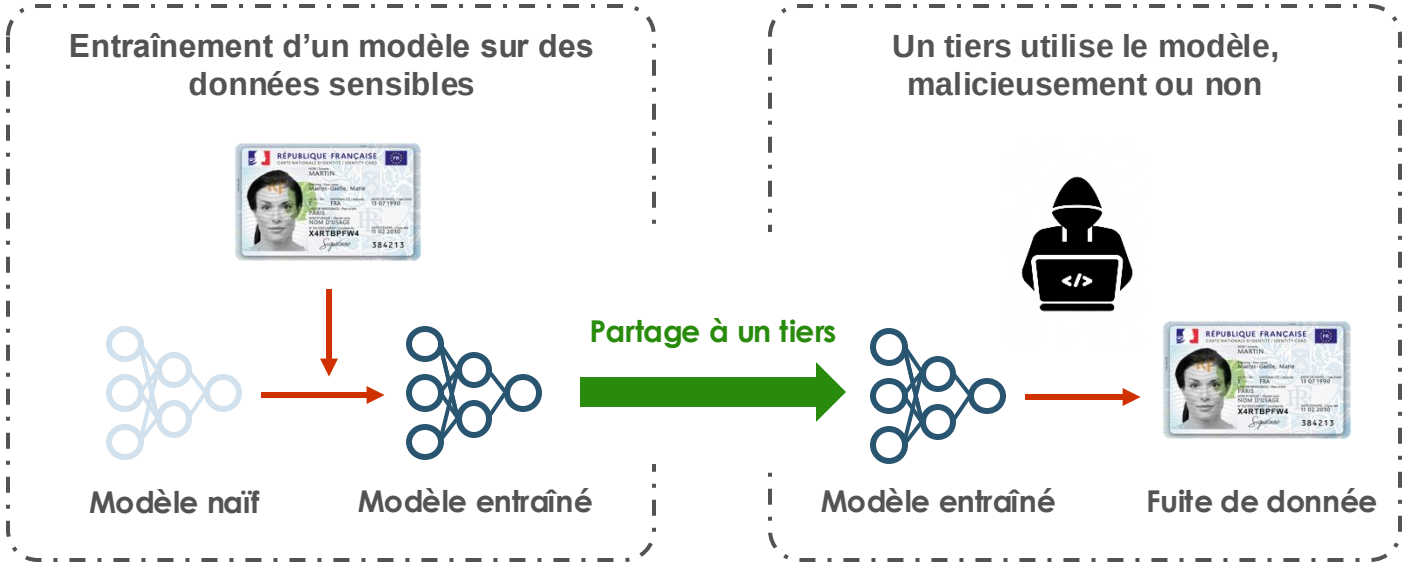
- Programme de **master sur deux années**
- Enseigné entièrement **en anglais**
- Orientation **professionnelle**
- Grade de **master**, pouvant être poursuivi en **doctorat**
- Enseignement dispensé par les professeurs du de **l'École Polytechnique** (DIX et CMAP), et les **entreprises partenaires**

Confidentialité des modèles de langage

- Mémorisation des données
- Les définitions de la mémorisation
- Prédire la mémorisation
- Méthode de résolution
- Comment ça marche ?
- Résultats

Mémorisation et modèles de langage

Les réseaux de neurone, notamment modèles de langage, mémorisent leurs données d'entraînement

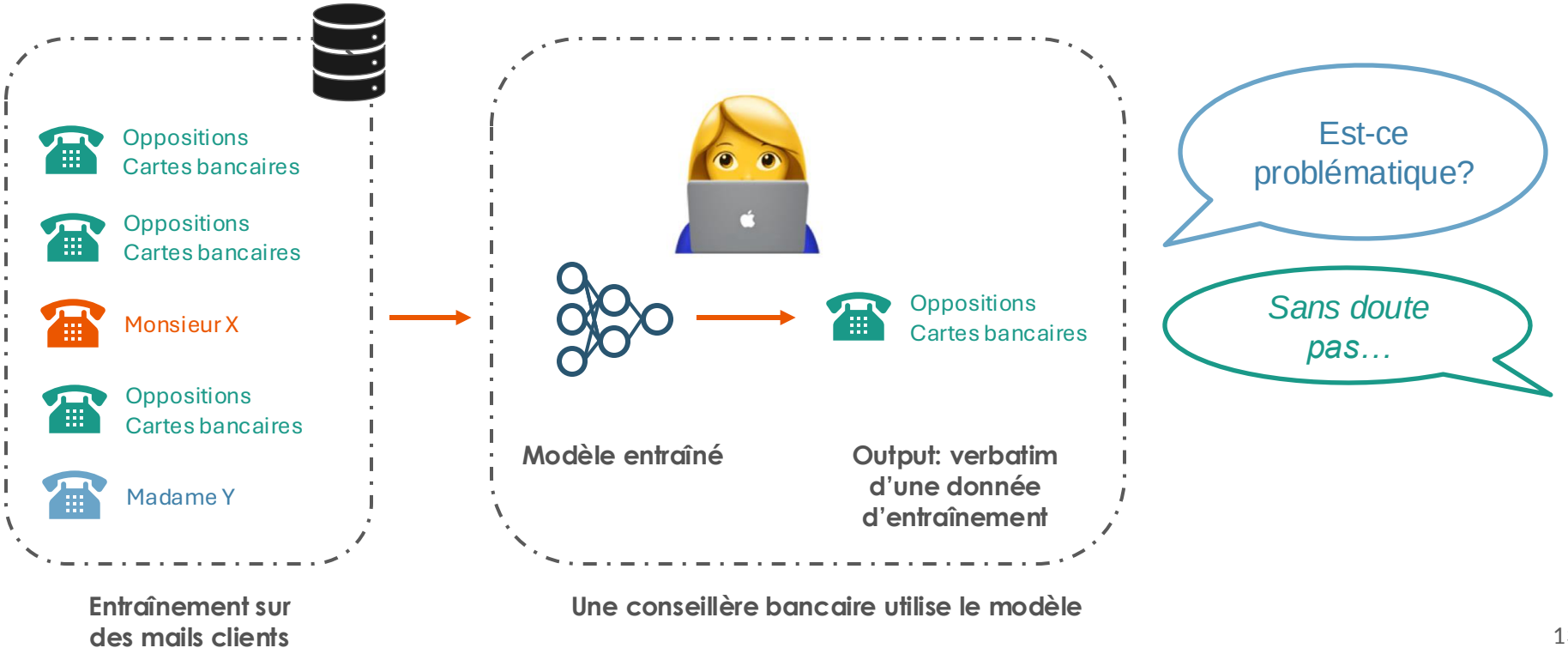


Cela peut arriver:

- **Par erreur**, par quiconque utilisant un modèle
- **Volontairement**, par un adversaire voulant extraire le plus de données possibles

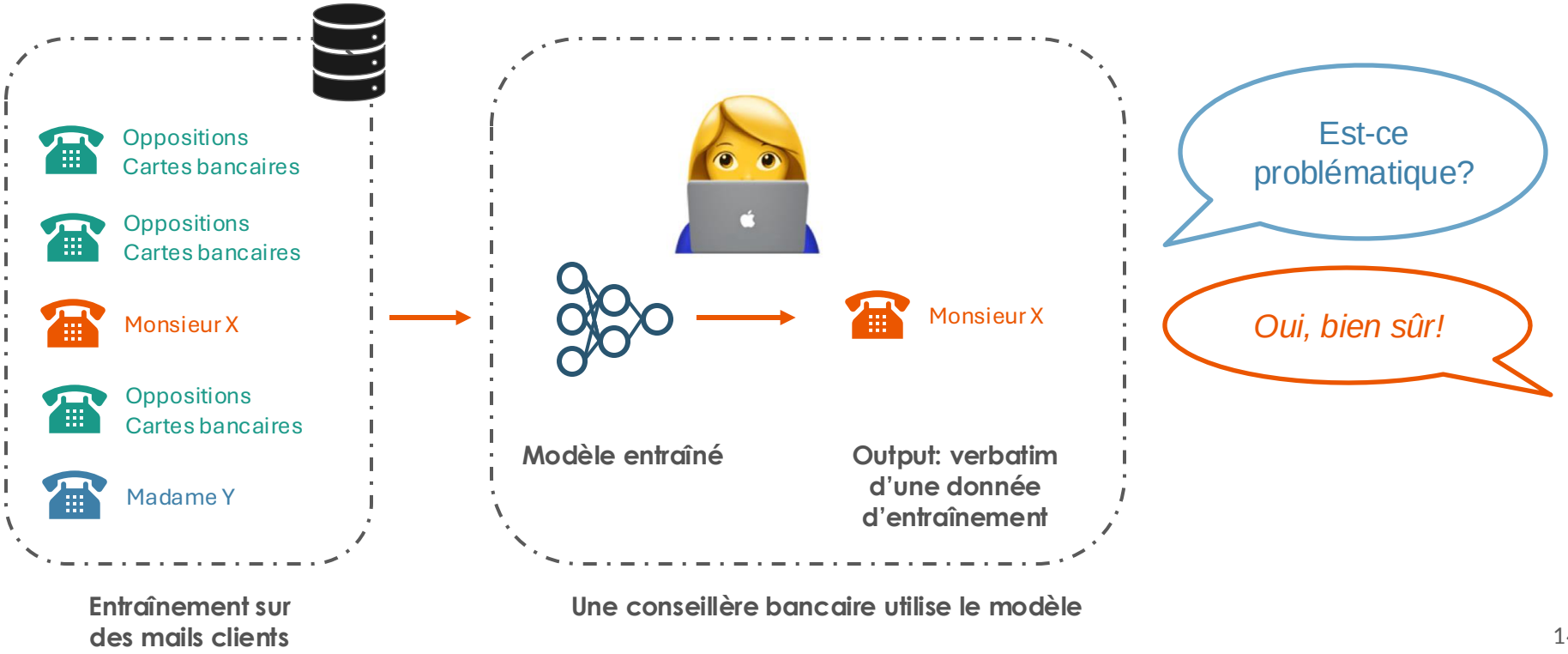
Mémorisation: un exemple

Le modèle renvoie le numéro de téléphone des oppositions carte bancaire, est-ce grave?



Mémorisation: un exemple

Le modèle renvoie le numéro de téléphone d'un client, est-ce grave?



Définir la mémorisation

La mémorisation des données d'entraînement est un concept complexe, avec plusieurs définitions

Extractibility

Est-il techniquement possible d'extraire la donnée en attaquant le modèle ?

Confidentialité différentielle

Une limite théorique de l'information qu'un adversaire peut obtenir

Inférence d'appartenance

Un adversaire peut-il savoir si ma donnée a été utilisée en entraînement ?

Mémorisation contrefactuelle

Quel est l'impact de chaque donnée sur les poids du modèle ?

➔ **Concrètement, comment savoir si mon modèle a mémorisé des données sensibles ?**

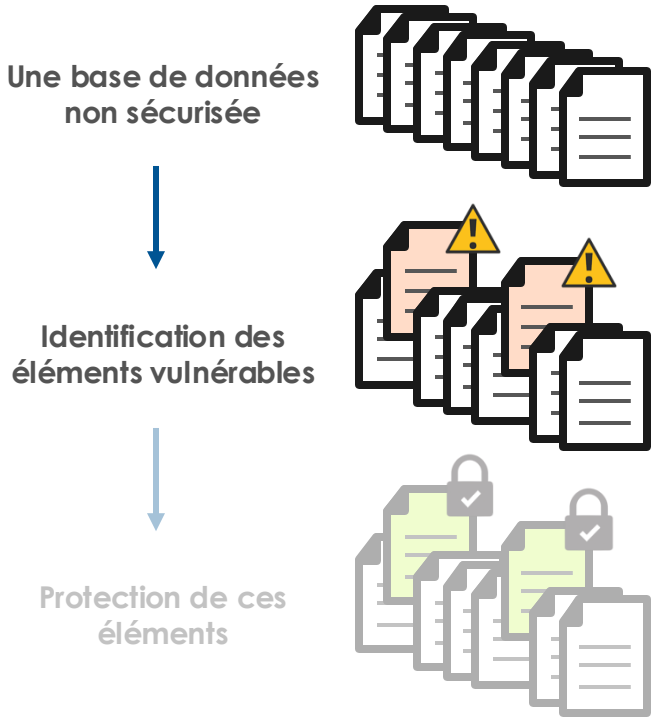
Prédire la mémorisation

Nous avons développé une approche pour auditer les modèles en cours d'entraînement

Scenario d'attaque:

- Des data scientist souhaite **auditer un modèle en cours de développement** et à moindre coût
- Exécuter des tests pour identifier les données vulnérables **avant qu'elles ne soient mémorisées.**
- Objectif long terme: protéger ces éléments efficacement et **à moindre coût.**

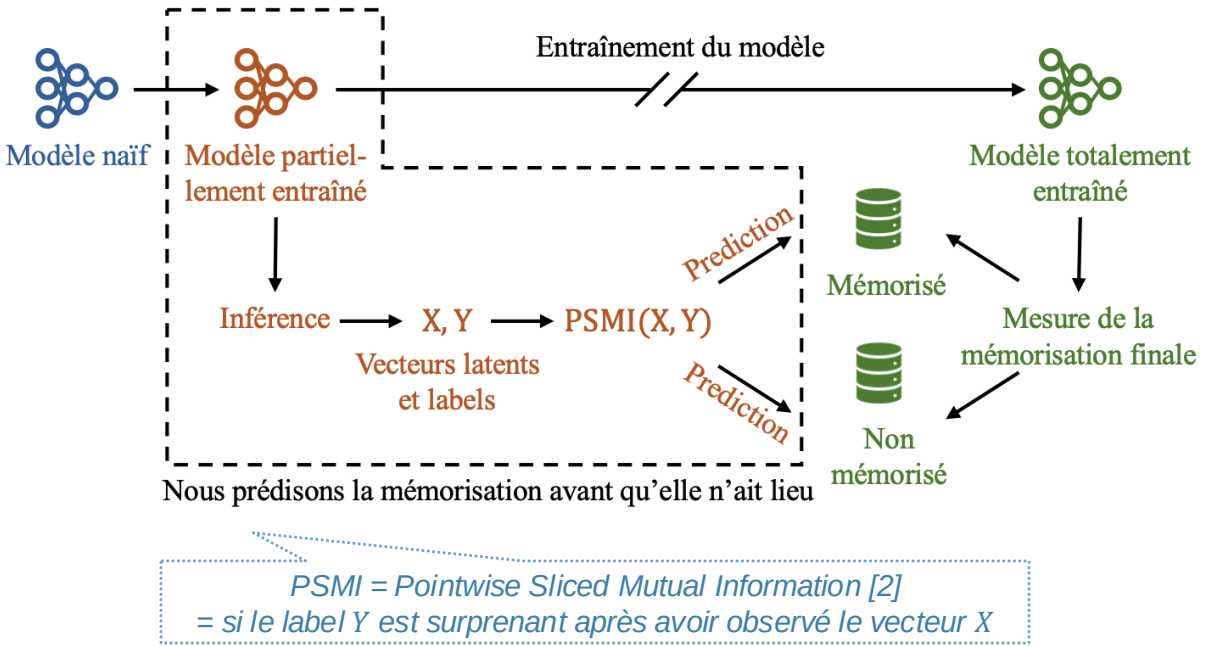
→ Ici: détection des données vulnérables. La protection post-détection sera étudiée dans des travaux futurs.



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

Méthode de résolution

Nous interrompons le modèle au début de l'entraînement pour prédire les données vulnérables.



PSMI = Pointwise Sliced Mutual Information [2]
= si le label Y est surprenant après avoir observé le vecteur X

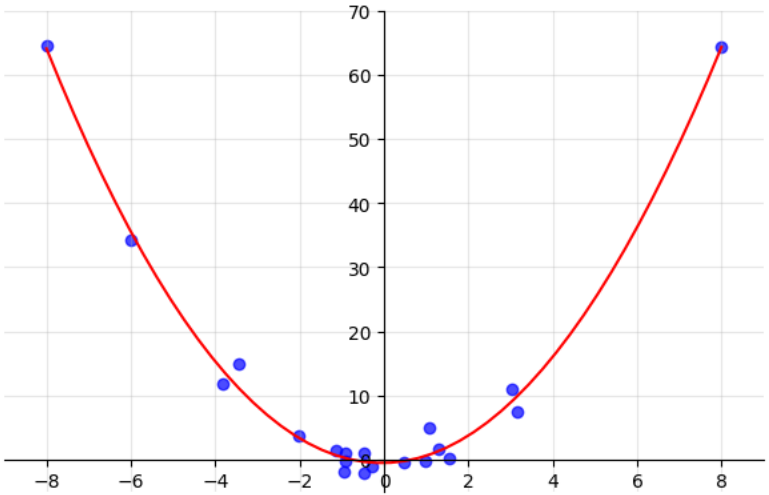
Points clefs:

- Nous prédisons la mémorisation **avant qu'elle n'ait lieu**
- Peu de calculs, **budget réaliste.**
- Étayée par des résultats théoriques, et **facilement adaptable** à n'importe quel problème de classification.

[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.

Comment ça marche?

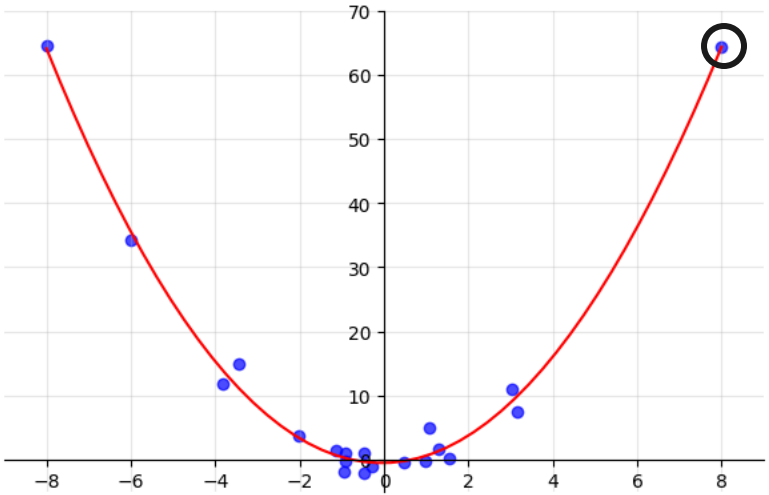
Les *outliers* ont un impact fort sur les poids d'un modèle, et sont donc plus fortement mémorisés.



Une régression polynomiale classique

Comment ça marche?

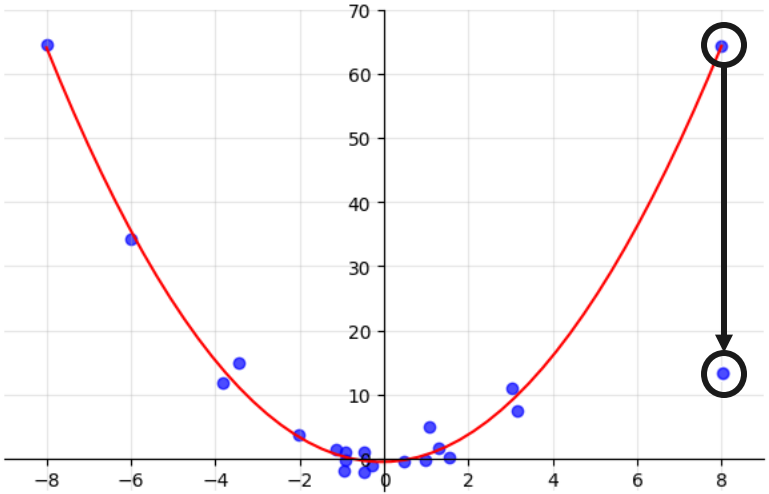
Les *outliers* ont un impact fort sur les poids d'un modèle, et sont donc plus fortement mémorisés.



Une régression polynomiale classique

Comment ça marche?

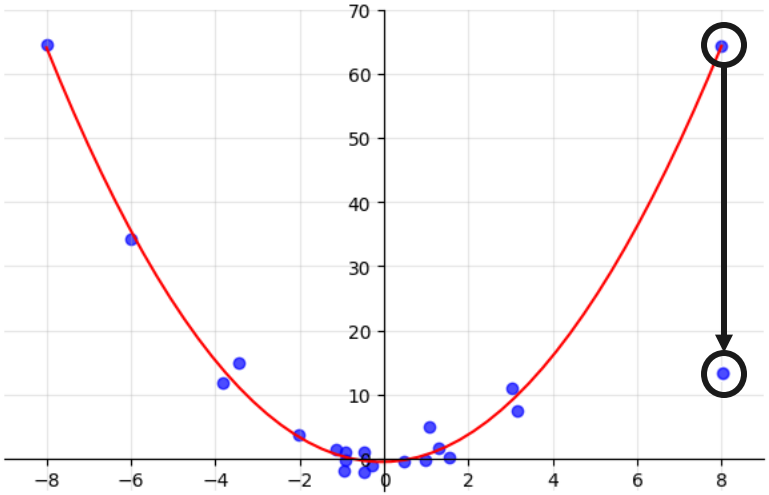
Les *outliers* ont un impact fort sur les poids d'un modèle, et sont donc plus fortement mémorisés.



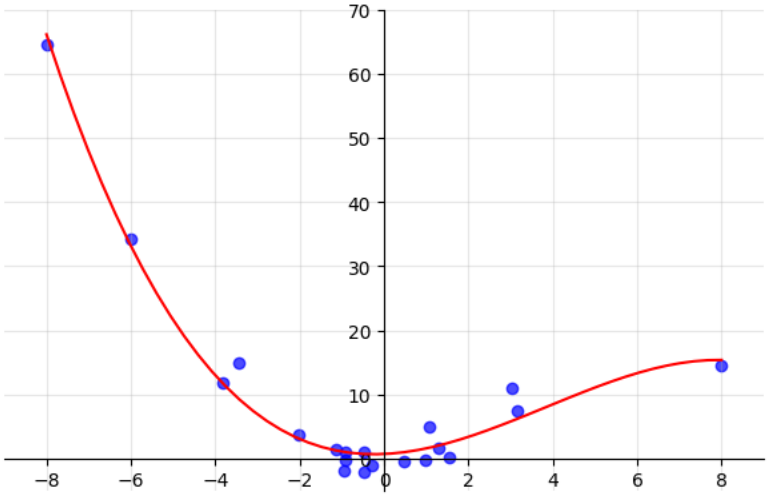
Une régression polynomiale classique

Comment ça marche?

Les *outliers* ont un impact fort sur les poids d'un modèle, et sont donc plus fortement mémorisés.



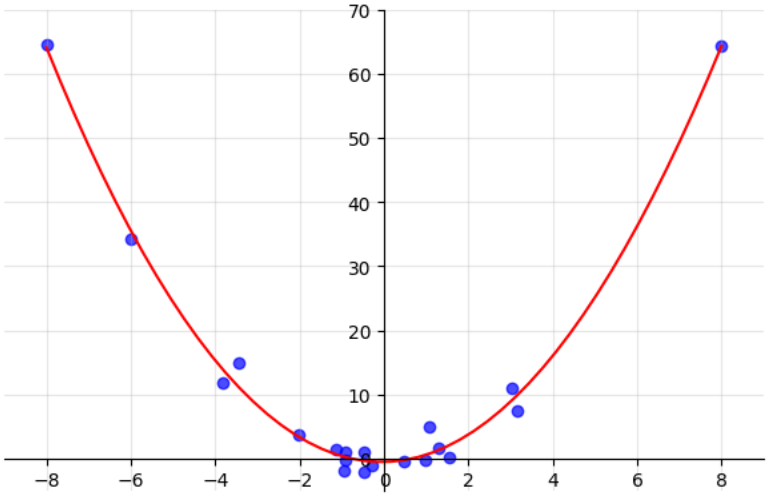
Une régression polynomiale classique



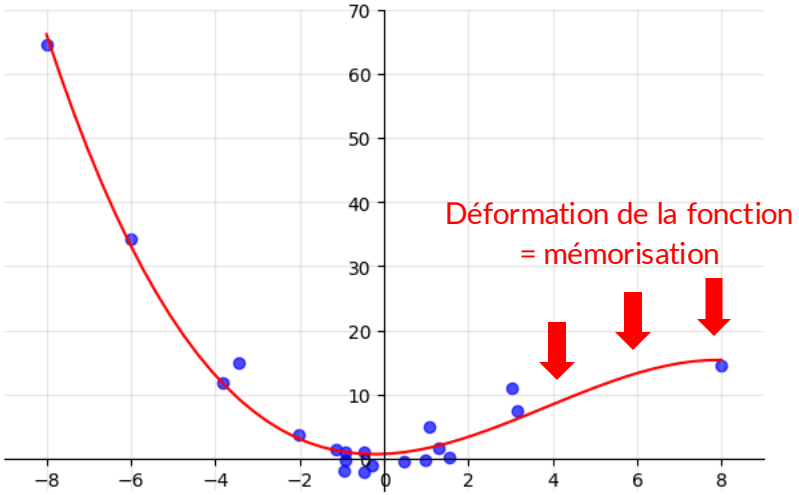
Une régression polynomiale avec *outlier*

Comment ça marche?

Les *outliers* ont un impact fort sur les poids d'un modèle, et sont donc plus fortement mémorisés.



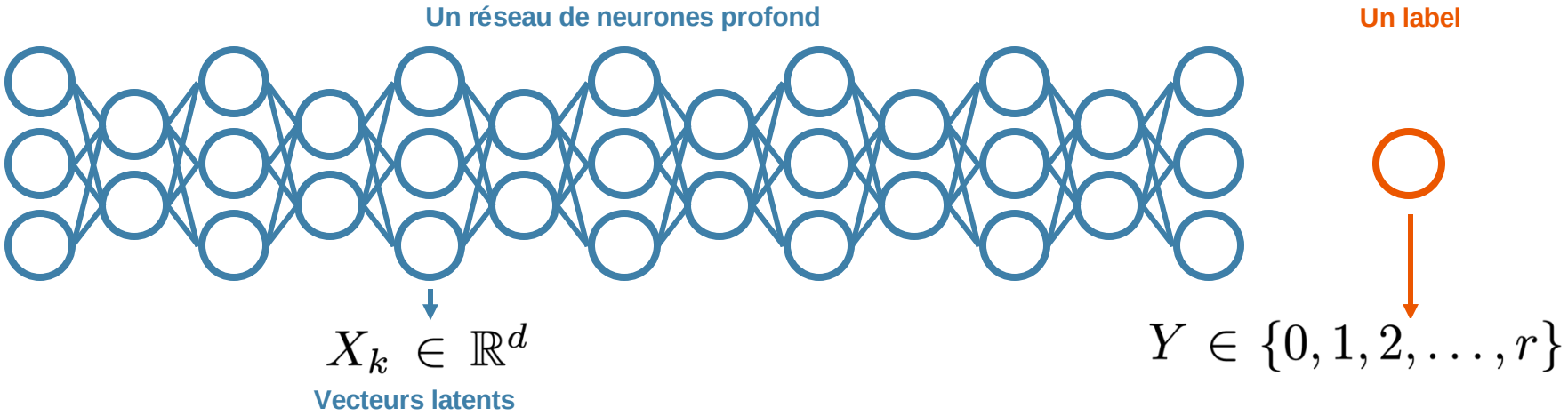
Une régression polynomiale classique



Une régression polynomiale avec *outlier*

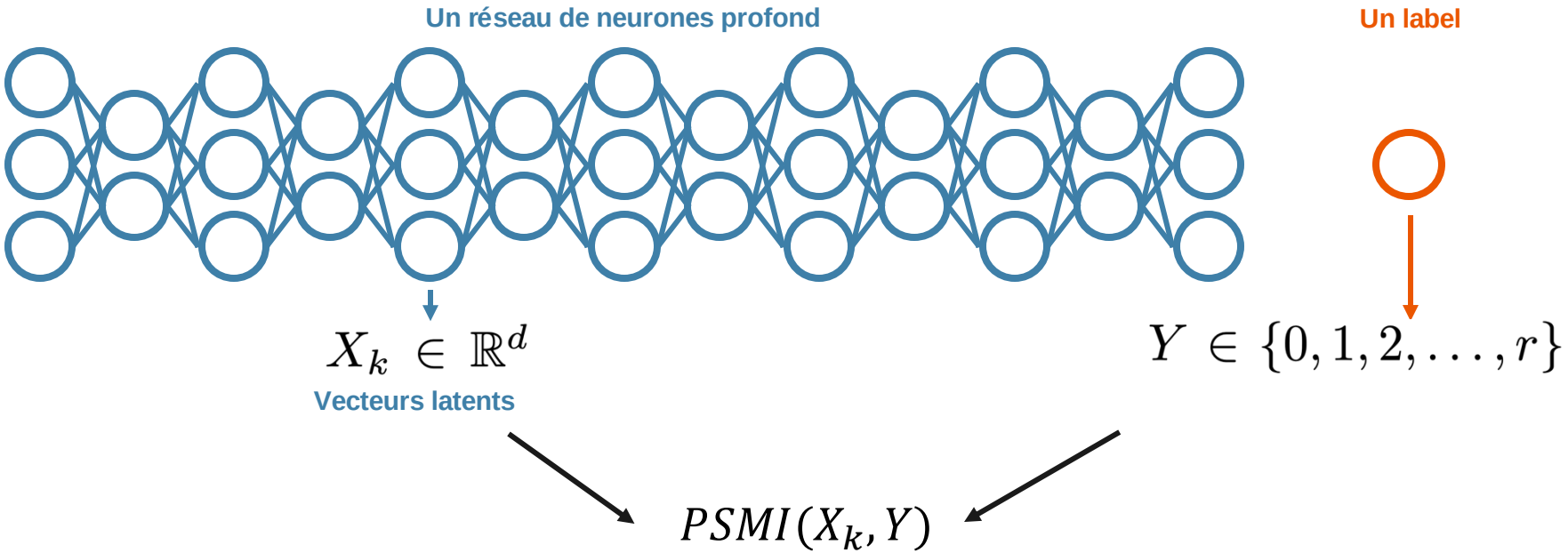
Et les réseaux de neurones?

Nous analysons les vecteurs latents du réseau pour détecter ces *outliers* qui risquent d'être mémorisés



Et les réseaux de neurones?

Nous analysons les vecteurs latents du réseau pour détecter ces *outliers* qui risquent d'être mémorisés

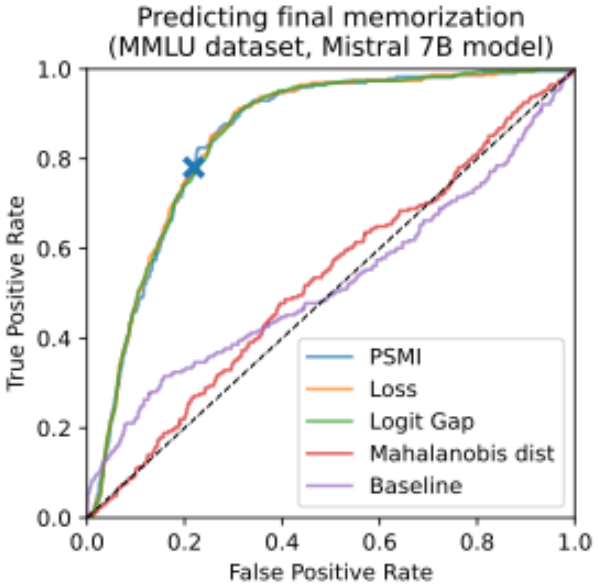


Information mutuelle entre les vecteurs latents et les labels, pour détecter ces outliers

De bons résultats empiriques

Nous avons validé notre approche sur cinq modèles de langages différents.

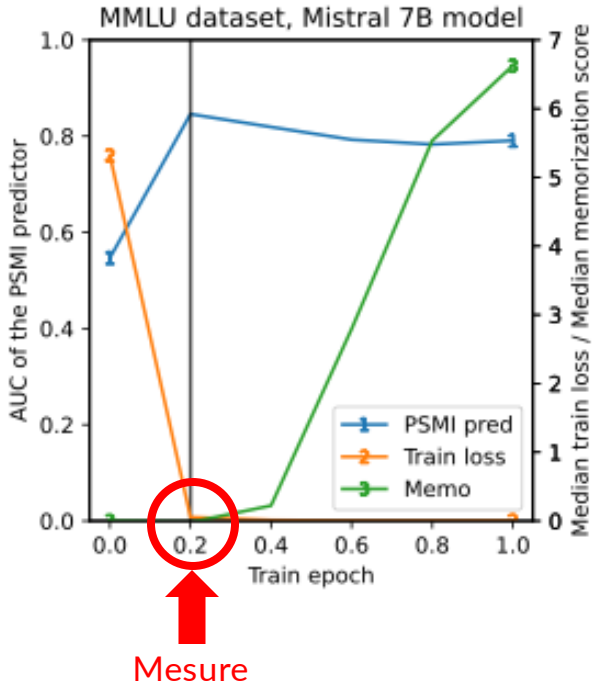
- Nous avons évalué notre approche avec **Gemma 7B, Mistral 7B et Llama 2 7B** fine-tunés pour la classification (MMLU, ARC, ETHICS).
- **Meilleures performances** que la seule baseline existante, pour **50 fois moins de calculs.**
- Ici: **FPR=21.9% et TPR=78.1%.**



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

Temporalité de la mesure

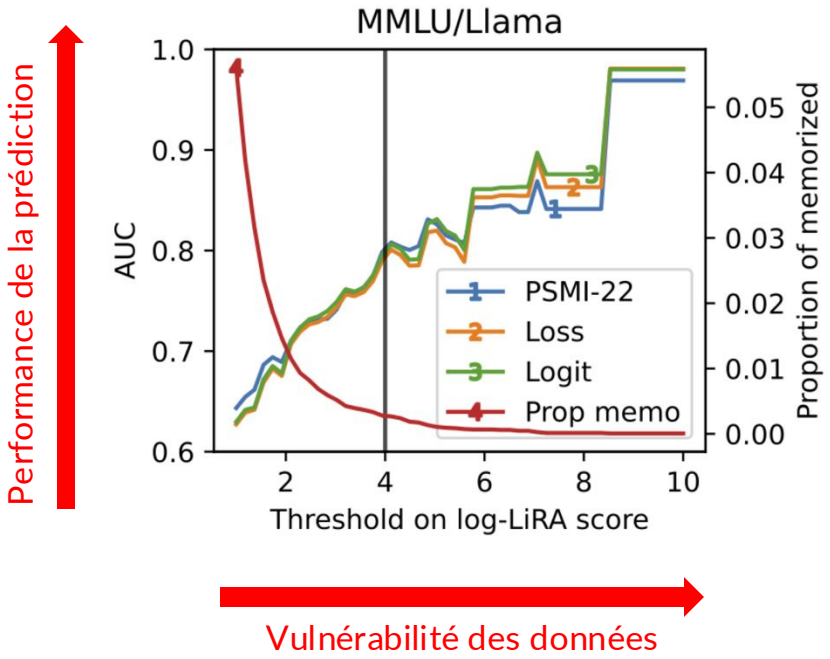
- Au moment, de la prédiction, les données ne sont **pas encore mémorisées.**
- Permet de protéger les données **sans recommencer** l'entraînement.



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

Seuils de vulnérabilité

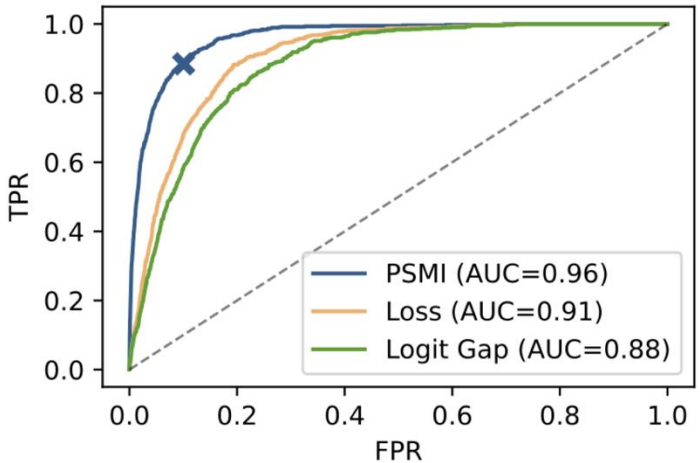
- Notre méthode est plus performante sur **les données les plus vulnérables.**
- Probabilité faible de manquer les données les plus vulnérables.



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

Une méthode adaptable

- Méthode utilisée avec succès sur un **modèle de vision**, sans adaptation nécessaire.
- **Adaptable** à d'autres scenarios de classification.



Mesure sur un WRN avec CIFAR-10, sans aucune adaptation

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

Et après?

Techniques de défense adaptatives

- **Protéger** les données vulnérables une fois détectées
- **Adapter** l'intensité de la protection à la vulnérabilité des données
- Améliorer le trade-off **confidentialité/performance**

Adaptation aux modèles génératifs

- Les expériences ont jusqu'ici été faites sur des modèles de classification
- **Adapter la méthode** aux modèles génératifs et à leurs risques accrus

Conclusion

- Détecter les *outliers* pour prédire les éléments **les plus susceptibles d'être mémorisés.**
- Une méthode **plus performante et 50 fois plus rapide** que la seule baseline existante.
- Méthode et hyperparamètres **adaptables** à d'autres scenarios de classification.

➔ *En savoir plus*

Merci

Financements

Ces travaux ont bénéficié du soutien financier de Crédit Agricole SA via la chaire de recherche « IA de Confiance et Responsable » avec l'École Polytechnique.

Ces travaux ont bénéficié de ressources de calcul en IA et de stockage au IDRIS au travers de l'allocation de ressources 2023-AD011014843 attribuée par GENCI sur la partition A100 du calculateur Jean Zay.