AYI NEDJIMI Consultants

Membership Inference Attacks LLM 2026 : Vol

16 mai 2026 • Mis à jour le 17 mai 2026 • 20 min de lecture • 3491 mots •

MIA permettent de prouver qu'un texte etait dans le training set d'un LLM. 0.65-0.85 sur Llama 4, implications RGPD majeures.

À RETENIR

A retenir — Membership Inference Attacks LLM

MIA determine si un texte etait dans le training set d'un LLM. AUC observe sur Llama 4 70B, 0.74 sur Claude 4.5 (Min-K%).

Familles d'attaque : *loss-based* (Shi et al., 2023), *Min-K%* (Shi et al., 2023) *Model*, *Zlib ratio*.

Implications RGPD : un sujet peut prouver que ses donnees personnelles o modele — droit a l'effacement applicable.

Defenses : DP-SGD (Abadi et al., 2016), redu n (deduplic gradient clipping, fine-tuning sans LoRA-Mer

> Litiges 2026 : *plaintiffs* utilisent MIA pour fonder leurs actions contre Open
> Stability AI, Meta. Charge probatoire elevee mais MIA y contribue.

Les **membership inference attacks** (*MIA*) sur LLM repondent a une question simp
juridiquement explosive : *"Ce texte etait-il dans le training set du modele ?"*. En 20
publient *Detecting Pretraining Data from Large Language Models* et demontrent q
statistique sur les log-probabilities, on distingue les samples membres (in-training
membres avec un AUC > 0.7 sur GPT-3, Llama, OPT. Trois ans plus tard, les MIA s
litiges juridiques (NYT vs OpenAI, Getty vs Stability) et de la conformite RGPD (dro
Cet article presente la mecanique mathematique, le code Python, les benchmarks
defenses (DP-SGD, deduplication). Pour les controleurs RGPD et les juristes speci
representent l'argumentaire technique le plus solide pour fonder le droit a l'efface
approfondit la **membership inference llm** sur Llama 4, Claude 4.5, GPT-5 et autres
avec implications concretes pour les fine-tunings d'entreprise.

1. Genese et etat de l'art

Le concept de MIA est introduit par Shokri et al. (2017) *Membership Inference Atta
Machine Learning Models, sur des classifiers d'images. Pour les LLM, la litterature
2022-2023 :

- Carlini et al. (2022) *The Privacy Risks of Memorization* — demonstration de me
  parfaite sur GPT-2.
- Carlini et al. (2023) *Quantifying Memorization Across Neural Language Models*
  de la memorisation.
- Shi et al. (2023) *Detecting Pretraining Data from L* k WIKIMIA,

Réponse sous 24h

Devis gratuit →