NOTE DE RECHERCHE

LLM et données personnelles : peut-on prouver qu'un modèle a mémorisé vos documents ?

Thomas Binant, Dream Big Consulting 29 octobre 2026 Version préliminaire

Statut de ce document. Note de recherche et synthèse de la littérature, signée par son auteur. Elle n'a pas fait l'objet d'une évaluation par les pairs et ne présente pas de résultat expérimental original : le protocole de la section 7 est une proposition, non testée à ce stade. Toutes les sources sont liées, pour que vous puissiez les vérifier. Titre complet : *Attribuer une donnée personnelle restituée par un modèle de langage à son document source : état de la littérature et protocole d'audit pour les actes authentiques*.

L'essentiel pour un décideur

- **Un modèle peut restituer un IBAN qu'il a vu.** C'est démontré, y compris sur des modèles de production.
- **Prouver qu'il vient d'un acte précis est peu fiable** sur un modèle de fondation : les méthodes font à peine mieux que le hasard.
- **Prouver qu'un fonds entier a été ingéré est possible**, avec un corpus témoin et un test statistique.
- **Demander au modèle sa source ne prouve rien** : sa réponse est une génération comme une autre.
- **La preuve se prépare avant** : des IBAN de contrôle fictifs rendent une ingestion attribuable après coup.

Résumé

Lorsqu'un grand modèle de langage (LLM) restitue une donnée personnelle précise, par exemple un identifiant bancaire (IBAN) figurant dans un acte authentique, deux questions distinctes se posent :

le modèle peut-il effectivement restituer cette donnée, et peut-on établir qu'elle provient de ce document particulier ? Cette note passe en revue la littérature en sécurité de l'apprentissage automatique sur l'extraction de données d'entraînement, l'inférence d'appartenance (*membership inference*), l'inférence de jeu de données (*dataset inference*), l'attribution par fonctions d'influence et les attaques contre les systèmes de génération augmentée par récupération (RAG). Nous montrons que la première question est tranchée positivement, y compris sur des modèles alignés en production, alors que la seconde admet des réponses très différentes selon le niveau de granularité visé (séquence, document, corpus) et le mode d'accès au modèle (boîte noire, probabilités de sortie, boîte blanche). Nous soulignons en particulier qu'interroger directement le modèle sur l'origine d'une sortie n'a aucun fondement dans la littérature. Nous proposons enfin un protocole d'audit à trois niveaux, incluant l'insertion préventive de canaris structurés (IBAN syntaxiquement valides mais fictifs) dans les fonds documentaires, afin de rendre attribuable, a posteriori et de manière statistiquement défendable, l'ingestion d'un fonds d'actes par un modèle tiers.

Mots-clés : mémorisation, extraction de données d'entraînement, inférence d'appartenance, attribution de données d'entraînement, RAG, données personnelles, actes authentiques, RGPD.

1. Introduction

Les grands modèles de langage sont entraînés sur des corpus de plusieurs milliers de milliards de tokens, collectés en grande partie par exploration du Web, complétés par des données licenciées ou produites en interne, et souvent affinés sur des documents propres à une organisation. Dans le secteur juridique et notarial, des documents à forte densité de données personnelles (actes de vente, actes de succession, procurations) peuvent ainsi se retrouver, volontairement ou non, dans les données d'entraînement d'un modèle, dans ses données d'affinage, ou dans la base documentaire d'un assistant fondé sur la génération augmentée par récupération (RAG).

La question pratique qui motive ce travail est la suivante : si un modèle restitue, en réponse à une sollicitation, un IBAN précis, est-il possible d'établir que cet IBAN provient d'un acte authentique identifié ? Cette question est au cœur de plusieurs enjeux : la qualification de la mémorisation au regard du RGPD [23], où le Comité européen de la protection des données a précisé en 2024 les conditions dans lesquelles un modèle d'IA peut ou non être considéré comme anonyme [24] ; la responsabilité de l'officier public dont le fonds aurait été exposé ; et, plus largement, l'auditabilité des modèles tiers.

Cette question unique recouvre en réalité trois problèmes scientifiques distincts, dont la maturité est très inégale : (i) l'*extraction*, c'est-à-dire la capacité à faire restituer la donnée ; (ii) l'*attribution*, c'est-à-dire la capacité à relier la donnée restituée à un document ou un corpus source ; (iii) le cas

particulier des architectures RAG, dans lesquelles le document n'est pas encodé dans les poids mais injecté dans le contexte.

2. Définitions et modèle de menace

Mémorisation. Plusieurs définitions coexistent dans la littérature [21]. Carlini *et al.* [1] introduisent la mémorisation *k*-eidétique : une séquence est mémorisée si elle est extractible par une sollicitation alors qu'elle n'apparaît qu'au plus *k* fois dans le corpus. Une distinction ultérieure [2], [25] sépare la mémorisation *découvrable* (la séquence est restituée lorsqu'on fournit son préfixe exact, ce qui suppose de connaître le corpus) de la mémorisation *extractible* (la séquence est restituée sans connaissance préalable du corpus). Seule la seconde correspond à un adversaire réaliste ne disposant pas de l'acte.

Fuite de données personnelles. Lukas *et al.* [3] formalisent trois jeux de sécurité : l'*extraction* (un attaquant non informé génère du texte et y repère des données personnelles), la *reconstruction* (un attaquant partiellement informé fournit le contexte d'une donnée masquée et demande au modèle de la compléter) et l'*inférence* (l'attaquant dispose d'un ensemble de candidats et demande au modèle de choisir). Le scénario « IBAN d'un acte » relève essentiellement de la reconstruction : l'attaquant connaît le vendeur, l'étude et la date, et sollicite le champ bancaire.

Inférence d'appartenance (MIA). Introduite pour les classifieurs par Shokri *et al.* [19], elle vise à déterminer si un enregistrement donné a fait partie du jeu d'entraînement. Pour les LLM, elle se décline au niveau de la séquence [18], du document [8] ou du jeu de données [9].

Attribution de données d'entraînement (TDA). Elle vise à quantifier la contribution de chaque exemple d'entraînement à une sortie donnée. La définition de référence est contrefactuelle (l'effet du retrait de l'exemple sur la sortie) mais exige un réentraînement par exemple, ce qui la rend inapplicable ; elle est approchée par les fonctions d'influence [11]–[13].

Modes d'accès. Nous distinguons l'accès *boîte noire texte* (seule la génération est observable), l'accès *boîte noire avec probabilités* (log-probabilités des tokens disponibles via l'API) et l'accès *boîte blanche* (poids, gradients et corpus). Cette distinction conditionne la faisabilité de chaque famille de méthodes.

3. Extraction : le modèle peut-il restituer l'IBAN ?

3.1 Mémorisation de séquences rares à forte entropie

Le résultat fondateur de Carlini *et al.* [1] sur GPT-2 établit que des séquences présentes une seule fois dans le corpus peuvent être extraites par un attaquant en boîte noire : parmi les 604 exemples mémorisés identifiés figurent des données de contact, des extraits de conversations et un mot de passe de 87 caractères apparaissant une unique fois. La vulnérabilité croît avec la taille du modèle. Ce point est décisif pour notre cas d'usage : un IBAN (27 caractères en France, forte entropie, structure régulière) présente précisément les caractéristiques des séquences que ces modèles mémorisent préférentiellement, et l'étude contrôlée de Meeus *et al.* [10] confirme causalement que les séquences à forte perplexité sont les plus vulnérables.

Carlini *et al.* [25] quantifient ensuite trois facteurs : la mémorisation croît log-linéairement avec la taille du modèle, avec le nombre de duplications de la séquence dans le corpus, et avec la longueur du contexte fourni. Un acte authentique est typiquement dédupliqué (une seule occurrence), ce qui réduit le risque, mais certains champs (coordonnées de l'étude, IBAN de séquestre de la Caisse des dépôts) se répètent sur des milliers d'actes et deviennent de ce fait fortement mémorisables.

3.2 Modèles alignés en production

Nasr *et al.* [2] étendent ces résultats aux modèles ouverts (Pythia, GPT-Neo), semi-ouverts (LLaMA, Falcon) et fermés (ChatGPT). Pour ce dernier, ils conçoivent une attaque de *divergence* qui fait sortir le modèle de son régime conversationnel et lui fait émettre des données d'entraînement à une fréquence 150 fois supérieure au régime normal. Des gigaoctets de données ont ainsi été extraits ; les auteurs concluent que les techniques d'alignement actuelles n'éliminent pas la mémorisation mais la masquent. Des travaux plus récents [26] généralisent ce principe en induisant délibérément un état de confusion du modèle.

3.3 Attaques ciblées sur les données personnelles

Lukas *et al.* [3] évaluent leurs trois attaques sur des modèles GPT-2 affinés dans trois domaines, dont la jurisprudence (*case law*). Leurs attaques extraient jusqu'à dix fois plus de séquences de données personnelles que les travaux antérieurs ; la confidentialité différentielle appliquée au niveau de la phrase réduit la fuite sans l'éliminer (environ 3 % des données restent extractibles) ; enfin, ils établissent un lien formel entre inférence d'appartenance au niveau de l'enregistrement et reconstruction, ce qui relie directement les sections 3 et 4 de cette note.

ProPILE [4] fournit l'outil le plus directement transposable. Il mesure la fuite par des sondes construites à partir de données déjà connues du sujet : une sonde « jumelle » (nom seul), « triplet »

(nom et une donnée) ou « quadruplet » (nom et deux données) pour faire compléter la donnée cible. Résultat central : l'ajout d'une donnée de contexte multiplie par cinq le taux de reconstruction exacte des adresses électroniques et par plus de deux celui des numéros de téléphone. Shao *et al.* [5] confirment que cette capacité d'*association*, distincte de la mémorisation verbatim, croît avec l'échelle du modèle. Transposé au notariat, une sonde de la forme « Acte reçu par Maître X le [date], vendeur Y, IBAN du vendeur : » constitue une attaque de reconstruction à contexte riche, et donc à fort rendement.

Bilan. L'extractibilité d'un IBAN mémorisé est démontrée, en boîte noire texte, sur des modèles de toutes tailles, y compris alignés. La probabilité de succès dépend principalement du nombre d'occurrences dans le corpus et de la richesse du contexte dont dispose l'attaquant.

4. Attribution : d'où vient la donnée restituée ?

L'extraction établit *que* le modèle restitue l'IBAN ; elle ne dit pas *d'où* il le tient. Il faut distinguer trois granularités, et le paysage scientifique change radicalement de l'une à l'autre.

4.1 Niveau document : l'inférence d'appartenance et sa remise en cause

Shi *et al.* [18] proposent Min-K% : une séquence est jugée membre du corpus si la moyenne des log-probabilités de ses *k* % tokens les moins probables dépasse un seuil. Meeus *et al.* [8] étendent l'approche au niveau du document entier : ils interrogent le modèle sur la probabilité de chaque vrai token, normalisent par la fréquence générale du token (pour ne pas confondre généralisation et mémorisation), agrègent et entraînent un classifieur d'appartenance. Antebi *et al.* [20] (Tag&Tab) améliorent le signal en ne considérant que les mots-clés sémantiquement saillants, ce qui, pour un acte, revient à peser davantage les noms propres, montants et identifiants.

Ces résultats ont toutefois été fortement relativisés. Duan *et al.* [7] évaluent cinq attaques sur les modèles Pythia (160 M à 12 B paramètres) avec un corpus d'entraînement connu, et constatent que les MIA font à peine mieux que le hasard dans la plupart des configurations. Deux causes sont identifiées : un corpus immense parcouru en peu d'époques (chaque séquence influence très peu les poids), et une frontière intrinsèquement floue entre membres et non-membres en langue naturelle (forte superposition de *n*-grammes). Les succès rapportés antérieurement s'expliquent par un biais de distribution : les non-membres étaient choisis dans une période temporelle différente des membres. Das *et al.* [6] vont plus loin : des attaques « aveugles », qui ignorent totalement le modèle cible et ne regardent que des caractéristiques du texte (date, vocabulaire), surpassent tous les résultats publiés des MIA sur les modèles de fondation. Chen *et al.* [27] confirment par une analyse statistique multi-perspectives que la plupart des méthodes ne surpassent pas significativement les baselines.

Conséquence pour notre cas. Établir, par MIA en boîte noire, qu'*un acte précis* a été vu par un modèle de fondation général est aujourd'hui scientifiquement peu fiable, et serait difficilement défendable devant une autorité ou un tribunal. La situation est différente pour un modèle *affiné* sur un petit corpus (plusieurs époques, forte influence de chaque document), configuration dans laquelle les MIA retrouvent leur efficacité [3], [7].

4.2 Niveau corpus : l'inférence de jeu de données

Maini *et al.* [9] reformulent le problème : plutôt que de tester l'appartenance d'une séquence, tester si un *ensemble* de documents a servi à l'entraînement. La méthode agrège les scores de plusieurs MIA, apprend par régression linéaire quelles MIA sont informatives pour la distribution considérée, puis applique un test statistique (test *t*) comparant le corpus suspect à un corpus de validation de même distribution jamais vu par le modèle. Sur les sous-ensembles de The Pile, la méthode distingue entraînement et validation avec des *p*-valeurs inférieures à 0,1, sans faux positif, et ne requiert qu'environ mille séquences.

Cette reformulation est exactement celle qui sert l'officier public : il ne cherche pas à prouver que l'acte n° 2024-1234 a été ingéré, mais que *son fonds d'actes* l'a été. Le prérequis, disposer d'un échantillon de même distribution certainement jamais vu par le modèle, est satisfaisable par la conservation d'actes postérieurs à la date de coupure du modèle, ou d'actes jamais numérisés.

4.3 Niveau sortie : l'attribution par fonctions d'influence

Seule la famille des méthodes d'attribution de données d'entraînement répond littéralement à la question « quel document a produit cette sortie ». Les fonctions d'influence [11] estiment l'effet d'une perturbation infinitésimale du poids d'un exemple d'entraînement via un produit scalaire des gradients corrigé par l'inverse de la Hessienne ; Grosse *et al.* [12] les ont appliquées à des modèles de 52 milliards de paramètres, et TRAK [13] ainsi que des travaux récents [28], [29] en réduisent le coût. Le résultat est un classement des exemples d'entraînement par influence sur la génération observée, c'est-à-dire précisément une attribution de source.

Trois limites rendent cette voie inaccessible à un auditeur externe : elle exige l'accès aux poids et aux gradients ; elle exige l'accès au corpus d'entraînement complet ; et son coût en calcul et en stockage reste prohibitif à l'échelle des corpus de préentraînement [29]. Fotouhi *et al.* [14] proposent une approximation en boîte noire par apprentissage en contexte, mais au niveau de la *source* (jeu de données) et sous une hypothèse de modularité que les auteurs eux-mêmes jugent peu réaliste. Cette attribution est donc un outil pour le fournisseur du modèle (conformité interne, réponse à une demande d'exercice de droits) et non pour le tiers lésé.

Question	Méthode de référence
Faire restituer l'IBAN	Extraction, divergence, sondes ProPILE [1]–[4
Prouver qu'un acte précis a été vu (modèle de fondation)	MIA niveau document [8], [18], [20]
Prouver qu'un acte précis a été vu (modèle affiné)	MIA et reconstruction [3]
Prouver qu'un fonds d'actes a été vu	Inférence de jeu de données [9]
Attribuer une sortie à un document	Fonctions d'influence, TRAK [11]–[13]
Idem en architecture RAG	Inférence d'appartenance et extraction fine [1
Demander sa source au modèle	Aucune

Tableau 1. Synthèse des questions, méthodes, conditions d'accès et niveau de preuve.

5. Le cas des architectures RAG

Lorsque l'acte n'est pas dans les poids mais dans une base documentaire interrogée à chaque requête, la fuite change de nature : le document est copié intégralement dans le contexte du modèle, et la frontière entre « généralisation » et « restitution » disparaît. Anderson *et al.* [17] montrent qu'il suffit de demander au système si un passage figure dans son contexte pour inférer l'appartenance d'un document à la base. Naseh *et al.* [15] rendent l'attaque furtive : des questions en langue naturelle, auxquelles seul le document cible permet de répondre, suffisent à inférer l'appartenance avec une trentaine de requêtes, pour un coût inférieur à 0,02 $ par document, tout en échappant aux détecteurs de requêtes adverses. Chen *et al.* [16] exploitent l'asymétrie de connaissance entre le système RAG et un LLM de référence sans base pour isoler, dans une réponse mixte, les fragments provenant de la base, avec plus de 90 % d'extraction en domaine unique.

Dans ce paradigme, l'attribution est intrinsèquement résolue : le système retrouve le document, le cite ou le paraphrase, et l'asymétrie avec un modèle de référence permet de localiser ce qui provient de la base. C'est donc le scénario le plus favorable à l'auditeur, et le plus défavorable à l'opérateur qui aurait indexé des actes sans contrôle d'accès.

6. Discussion

6.1 L'interrogation directe du modèle n'est pas une méthode

Aucune des méthodes recensées ne consiste à demander au modèle d'où provient une information. Les signaux exploités sont des probabilités de tokens, des gradients ou des divergences entre

systèmes, jamais une auto-déclaration. Un LLM ne dispose pas d'un accès introspectif à son corpus d'entraînement : une réponse du type « cet IBAN provient de l'acte reçu par Maître X » serait générée comme n'importe quelle autre complétion plausible et doit être traitée comme une confabulation. Pire, elle peut constituer elle-même une fuite supplémentaire si le nom de l'étude ou du client a été mémorisé en association avec l'IBAN [4], [5].

6.2 Validité probante

La littérature sur les biais d'évaluation des MIA [6], [7], [9] impose une discipline : tout audit doit comporter un corpus témoin de même distribution, certainement non vu par le modèle, et rapporter des taux de vrais positifs à faible taux de faux positifs plutôt qu'une aire sous la courbe moyenne. Des travaux récents sur les limites probatoires de l'inférence d'appartenance pour l'audit de droits d'auteur [30] vont dans le même sens.

6.3 Lecture juridique

L'avis 28/2024 du CEPD [24] conditionne l'anonymat d'un modèle à une probabilité insignifiante, pour tous les moyens raisonnablement susceptibles d'être utilisés, d'extraire des données personnelles ou de les obtenir par des requêtes. Les résultats des sections 3 et 5 montrent que, pour des identifiants structurés à forte entropie présents dans un corpus d'affinage ou une base RAG, cette probabilité n'est généralement pas insignifiante. En revanche, l'imputabilité à un document précis, nécessaire pour établir par exemple une violation de données au sens de l'article 33 du RGPD à la charge d'un responsable identifié, ne peut être fondée sur une MIA au niveau document contre un modèle de fondation, et devra s'appuyer sur l'inférence au niveau corpus ou sur des preuves préconstituées.

7. Protocole d'audit proposé

Nous proposons un protocole en trois niveaux, du moins au plus exigeant en accès, applicable par un officier public, un éditeur de logiciel notarial ou une autorité.

Niveau A : extraction (boîte noire texte). Construire, à partir d'actes réels tirés au sort, des sondes de reconstruction à la ProPILE : contexte croissant (étude seule ; étude et date ; étude, date et parties) et champ cible masqué (IBAN, numéro de compte de séquestre, adresse). Mesurer le taux de reconstruction exacte et partielle (code banque, code guichet) en fonction de la richesse du contexte, et le comparer à un modèle de référence dont on sait qu'il n'a pas vu le fonds. Un écart significatif établit l'extractibilité, non l'attribution.

Niveau B : inférence de corpus (boîte noire avec log-probabilités). Appliquer la procédure de Maini *et al.* [9] : échantillon suspect d'au moins mille séquences issues du fonds ; échantillon témoin de même distribution (actes postérieurs à la date de coupure du modèle, ou jamais numérisés) ; agrégation de MIA incluant Min-K% [18] et une variante par mots-clés [20] ; test statistique avec correction. Rapporter la *p*-valeur et le taux de vrais positifs à 1 % de faux positifs. Ce niveau produit la preuve exploitable : « ce fonds a été ingéré ».

Niveau C : canaris structurés (preuve préconstituée). Inspiré des *copyright traps* [10], dont les résultats montrent que des séquences longues et répétées sont détectables même dans des modèles qui mémorisent peu, ce niveau consiste à insérer, en amont de toute exposition, des IBAN de contrôle : syntaxiquement valides (clé de contrôle correcte), rattachés à des codes établissement réservés ou fictifs, uniques par étude ou par lot, et enregistrés dans un registre horodaté. Un IBAN canari cumule les propriétés que la littérature identifie comme favorables à la mémorisation : forte entropie [1], [10] et, si on le répète dans les mentions de séquestre, forte duplication [25]. Sa restitution ultérieure par un modèle tiers, obtenue au niveau A, constitue une attribution non ambiguë, puisque la chaîne n'existe nulle part ailleurs. Les mêmes canaris alimentent le niveau B en séquences membres certaines.

Ce protocole présente deux limites. D'une part, les canaris ne protègent que les documents produits après leur mise en place ; le stock d'actes existant reste justiciable des seuls niveaux A et B. D'autre part, l'insertion d'une mention fictive dans un acte authentique soulève une question de forme qui dépasse cette note ; une alternative est d'insérer les canaris dans les métadonnées, annexes ou versions de travail, qui sont précisément celles le plus souvent exposées.

8. Conclusion

La capacité d'un LLM à restituer une donnée personnelle rare, telle qu'un IBAN figurant dans un acte authentique, est solidement établie, y compris sur des modèles de production alignés. La capacité à prouver que cette donnée provient d'un document donné dépend en revanche du niveau de granularité et du mode d'accès : peu fiable au niveau du document isolé contre un modèle de fondation, statistiquement robuste au niveau du corpus, et réservée au fournisseur du modèle au niveau de la sortie individuelle. Dans les architectures RAG, le problème est largement résolu dans les deux sens. Enfin, interroger le modèle sur l'origine de sa sortie n'a pas de fondement scientifique. Nous recommandons aux professions produisant des fonds documentaires sensibles de ne pas attendre un hypothétique outil d'attribution a posteriori, mais de préconstituer la preuve par des canaris structurés et de conserver des corpus témoins permettant l'inférence au niveau corpus.

Références

1. N. Carlini, F. Tramèr, E. Wallace *et al.*, « Extracting Training Data from Large Language Models », *Proc. 30th USENIX Security Symposium*, 2021. arXiv:2012.07805
2. M. Nasr, N. Carlini, J. Hayase *et al.*, « Scalable Extraction of Training Data from (Production) Language Models », *Proc. ICLR*, 2025. arXiv:2311.17035
3. N. Lukas, A. Salem, R. Sim, S. Tople, L. Wutschitz et S. Zanella-Béguelin, « Analyzing Leakage of Personally Identifiable Information in Language Models », *Proc. IEEE Symposium on Security and Privacy*, 2023. arXiv:2302.00539
4. S. Kim, S. Yun, H. Lee, M. Gubri, S. Yoon et S. J. Oh, « ProPILE: Probing Privacy Leakage in Large Language Models », *NeurIPS*, 2023. arXiv:2307.01881
5. H. Shao, J. Huang, S. Zheng et K. C.-C. Chang, « Quantifying Association Capabilities of Large Language Models and Its Implications on Privacy Leakage », *Findings of EACL*, 2024. arXiv:2305.12707
6. D. Das, J. Zhang et F. Tramèr, « Blind Baselines Beat Membership Inference Attacks for Foundation Models », 2024. arXiv:2406.16201
7. M. Duan, A. Suri, N. Mireshghallah *et al.*, « Do Membership Inference Attacks Work on Large Language Models? », *Proc. COLM*, 2024. arXiv:2402.07841
8. M. Meeus, S. Jain, M. Rei et Y.-A. de Montjoye, « Did the Neurons Read your Book? Document-level Membership Inference for Large Language Models », *Proc. 33rd USENIX Security Symposium*, 2024. arXiv:2310.15007
9. P. Maini, H. Jia, N. Papernot et A. Dziedzic, « LLM Dataset Inference: Did you train on my dataset? », *NeurIPS*, 2024. arXiv:2406.06443
10. M. Meeus, I. Shilov, M. Faysse et Y.-A. de Montjoye, « Copyright Traps for Large Language Models », *Proc. ICML*, 2024. arXiv:2402.09363
11. P. W. Koh et P. Liang, « Understanding Black-box Predictions via Influence Functions », *Proc. ICML*, 2017. arXiv:1703.04730
12. R. Grosse, J. Bae, C. Anil *et al.*, « Studying Large Language Model Generalization with Influence Functions », 2023. arXiv:2308.03296
13. S. M. Park, K. Georgiev, A. Ilyas, G. Leclerc et A. Madry, « TRAK: Attributing Model Behavior at Scale », *Proc. ICML*, 2023. arXiv:2303.14186
14. M. Fotouhi, M. T. Bahadori, O. Feyisetan, P. Arabshahi et D. Heckerman, « Fast Training Dataset Attribution via In-Context Learning », 2024. arXiv:2408.11852

15. A. Naseh, Y. Peng, A. Suri, H. Chaudhari, A. Oprea et A. Houmansadr, « Riddle Me This! Stealthy Membership Inference for Retrieval-Augmented Generation », *Proc. ACM CCS*, 2025. arXiv:2502.00306

16. Y. Chen, Y. Wang, H. Zhang et T. Gu, « Fine-Grained Privacy Extraction from Retrieval-Augmented Generation Systems via Knowledge Asymmetry Exploitation », 2025. arXiv:2507.23229

17. M. Anderson, G. Amit et A. Goldsteen, « Is My Data in Your Retrieval Database? Membership Inference Attacks Against Retrieval Augmented Generation », 2024. arXiv:2405.20446

18. W. Shi, A. Ajith, M. Xia *et al.*, « Detecting Pretraining Data from Large Language Models », *Proc. ICLR*, 2024. arXiv:2310.16789

19. R. Shokri, M. Stronati, C. Song et V. Shmatikov, « Membership Inference Attacks Against Machine Learning Models », *Proc. IEEE Symposium on Security and Privacy*, 2017. arXiv:1610.05820

20. S. Antebi, E. Habler, A. Shabtai et Y. Elovici, « Tag&Tab: Pretraining Data Detection in Large Language Models Using Keyword-Based Membership Inference Attack », *Findings of EMNLP*, 2025. arXiv:2501.08454

21. S. Ishihara, « Training Data Extraction From Pre-trained Language Models: A Survey », 2023. arXiv:2305.16157

22. A. Duarte, X. Zhao, A. Oliveira et L. Li, « DE-COP: Detecting Copyrighted Content in Language Models Training Data », *Proc. ICML*, 2024. arXiv:2402.09910

23. Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 (RGPD), *JOUE* L 119, 4 mai 2016. Texte officiel

24. Comité européen de la protection des données, *Avis 28/2024 sur certains aspects relatifs à la protection des données dans le contexte du traitement de données à caractère personnel aux fins de modèles d'IA*, adopté le 17 décembre 2024. Texte officiel

25. N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramèr et C. Zhang, « Quantifying Memorization Across Neural Language Models », *Proc. ICLR*, 2023. arXiv:2202.07646

26. M. Ko, N. R. Billa, A. Nguyen, C. Fleming, M. Jin et R. Jia, « Retracing the Past: LLMs Emit Training Data When They Get Lost », 2025. arXiv:2511.05518

27. B. Chen, N. Han et Y. Miyao, « A Statistical and Multi-Perspective Revisiting of the Membership Inference Attack in Large Language Models », *Proc. ACL*, 2025. arXiv:2412.13475

28. F. Berdoz, L. A. Lanzendörfer, K. Bayraktar et R. Wattenhofer, « Data Attribution in Large Language Models via Bidirectional Gradient Optimization », 2026. arXiv:2606.04928

29. D. Kachler, D. Sileo et P. Denis, « Influcoder: Distilling Decoders' Gradient Influence Rankings into an Encoder for Data Attribution », 2026. arXiv:2606.13668

30. M. B. Ertan, E. Böge, M. Chen, K. Mahmood et M. van Dijk, « On the Evidentiary Limits of Membership Inference for Copyright Auditing », 2026. arXiv:2601.12937