Appel à manifestation d'intérêt (AMI) : Test de la librairie PANAME (*Privacy Auditing of AI Models*)

La CNIL, l'ANSSI, le PEReN et le projet IPoP du PEPR (programmes et équipements prioritaires de recherche) Cybersécurité piloté par Inria ont lancé en juin 2025 le projet PANAME (*Privacy Auditing of AI Models*). Ce projet vise à développer un outil pour auditer la confidentialité des modèles d'IA, à travers le développement d'**une librairie permettant de réaliser des tests d'extraction et/ou de réidentification de données sur les modèles d'IA.**

Dans le cadre de ce projet, un Appel à Manifestation d'Intérêt (AMI) est lancé auprès des acteurs de l'écosystème souhaitant participer à la phase de test de cette librairie. Les retours d'expérience et la mise en œuvre de ces tests permettront d'enrichir les fonctionnalités et notamment d'assurer l'adéquation de la librairie aux besoins opérationnels d'évaluation de conformité RGPD.

Cet AMI s'adresse aux entreprises, startups, laboratoires de recherche et institutions utilisant ou développant des modèles d'IA, et souhaitant évaluer la robustesse de leurs systèmes face aux risques de mémorisation de données d'entraînement.

Cet AMI est **ouvert du 26 Février 2026 jusqu'au 28 Mars 2026**. Les réponses peuvent être déposées à compter de la date de publication de cet AMI. Les dossiers doivent être adressés exclusivement sous forme électronique à l'adresse mail : paname-test@inria.fr et respecter le format dédié « formulaire de réponse à l'AMI » disponible en annexe 1 du présent appel.

Suite à cet AMI, un atelier sera organisé en hybride par la CNIL, l'ANSSI, le PEReN et Inria le 13 avril 2026 après-midi à la CNIL, pour convier les acteurs ayant répondu à cet AMI avec pour objectifs de :

1) Présenter la librairie PANAME et les types de tests attendus.
2) Faciliter la prise en main de la libraire PANAME et répondre aux questions des acteurs.
3) Lancer officiellement la phase de tests de la libraire PANAME et partager le rétroplanning.

1. Contexte de l'appel à manifestation d'intérêt

1.1. Modèles d'IA et protection des données personnelles

Depuis plus d'une décennie, il est établi scientifiquement que le processus d'entraînement d'un modèle d'IA peut mener à la possibilité d'en extraire, a posteriori, des informations concernant les données présentes dans le jeu d'entraînement. Cette extraction peut se mener par des techniques statistiques (voir Annexe 2) au niveau du modèle par un accès complet ou partiel à celui-ci (fonction de coût, gradient, activations…), ou bien, dans le cas de l'IA générative, en requêtant directement le modèle par prompt. Si ces techniques d'extraction ont été développées bien avant l'IA générative, le fait de pouvoir extraire des données d'entraînement en langage naturel de ces systèmes utilisés par une part importante de la population, a porté cet enjeu à l'attention du grand public.

Dès lors que les données d'entraînement sont soumises à une réglementation particulière, comme c'est le cas par exemple pour des données à caractère personnel, des données protégées par le droit d'auteur, ou bien par le secret professionnel, la question de l'application de la réglementation au modèle entraîné se pose. En particulier, l'avis adopté par le comité européen de la protection des données (CEPD) en décembre 2024 rappelle que le RGPD s'applique, dans de nombreux cas, aux modèles d'IA entraînés sur des données personnelles, en raison de leurs capacités de mémorisation.

Il précise également qu'il est très souvent nécessaire de démontrer dans une analyse qu'un modèle entraîné sur des données personnelles résiste à des tests d'extraction d'informations sur les données d'entraînement pour conclure à son caractère anonyme, condition permettant de sortir son utilisation du champ d'application du RGPD.

1.2. Des ressources encore peu disponibles

Depuis une dizaine d'années, les chercheurs multiplient les travaux sur les méthodes d'extraction. Toutefois, leur mise en œuvre est souvent faite à un niveau expérimental, pour des publications scientifiques. Plusieurs obstacles ont ainsi été identifiés pour leur appropriation par les industriels :

- **Une littérature académique éparse et foisonnante :** se repérer dans les ressources sur le sujet peut demander du temps et des compétences techniques élevées, notamment pour les plus petits acteurs ;
- **Des implémentations pas toujours adaptées à un contexte industriel :** même lorsqu'elles sont disponibles en source ouverte (open source), ces techniques nécessitent un important travail de développement et d'intégration pour être utilisées en production ;
- **Un manque de standardisation :** aucun cadre unifié n'existe actuellement pour formaliser le codage des tests de confidentialité.

1.3. Le projet PANAME comme outil pour répondre à ces enjeux

Afin de répondre aux enjeux d'évaluation de conformité RGPD et de lever les freins identifiés, la CNIL, l'ANSSI, le PEReN et le projet IPoP du PEPR (programmes et équipements prioritaires de recherche) Cybersécurité piloté par Inria ont lancé en juin 2025 le projet PANAME (*Privacy Auditing of AI Models*).

Ce projet de 18 mois permettra de développer une bibliothèque logicielle disponible toute ou partie en source ouverte, destinée à unifier la façon dont la confidentialité des modèles est évaluée.

L'objectif de l'outil est de permettre une mise en œuvre efficace et à moindre coût de certains tests techniques d'extraction d'informations sur les données d'entraînement que les acteurs de l'écosystème IA sont susceptibles de devoir réaliser pour évaluer le statut d'un modèle d'IA au regard du RGPD.

Après des premiers mois de travail autour des spécifications techniques et des premiers développements sur la libraire, la CNIL, Le PEReN, l'ANSSI et Inria souhaitent lancer une phase de tests avec des administrations et des industriels afin de s'assurer que le développement de l'outil se fait en cohérence avec leur contexte d'utilisation.

Le présent appel à manifestation d'intérêt vise à identifier les contributeurs de cette phase de test de la librairie PANAME.

2. Description et attendus de l'appel à manifestation d'intérêts « Tests de la librairie PANAME »

- **Le présent AMI est ouvert du 26 février 2026 jusqu'au 28 Mars 2026.** Les réponses doivent être adressées exclusivement sous forme électronique à l'adresse mail : paname-test@inria.fr
- **Les acteurs répondant à cet AMI sont tenus de respecter le format dédié « formulaire de réponse à l'AMI » disponible en annexe 1 du présent AMI.**
- **Cet AMI n'est pas doté financièrement :** la participation à cette phase de tests s'effectue sur la base du volontariat et ne donne lieu à aucune rémunération financière par les membres du projet PANAME. Les acteurs industriels et institutionnels sélectionnés contribueront ainsi à l'amélioration collective de la librairie PANAME et pourront également tester la confidentialité de leurs systèmes d'IA avec celle-ci.
- **Cet AMI constitue la phase 0 des tests de la librairie PANAME. Le cadrage global de la phase de tests de la librairie PANAME est donné ci-après.**

2.1. Cadrage global de la phase de tests de la librairie PANAME

La phase de test de la librairie PANAME sera découpée en trois phases :

- **Phase 0 - Appel à Manifestation d'intérêts pour identifier les testeurs de la libraire PANAME :** cette phase objet du présent document vise à identifier les acteurs intéressés pour tester la librairie PANAME ;
- **Phase 1 - 1er Tests de la librairie PANAME avec l'ensemble de l'écosystème :** suite à l'identification des

volontaires en phase 0, cette phase permet à l'ensemble de l'écosystème de tester la librairie PANAME et de contribuer à son amélioration par voie de retour sur les tests ;

- **Phase 2 (en opportunité) - Tests approfondis de la librairie de PANAME avec les testeurs référents :** suite à la phase 1 et aux premières restitutions de tests, la CNIL, le PEReN, l'ANSSI et Inria pourront éventuellement sélectionner des testeurs référents ayant été le plus impliqués dans la 1ère phase de test afin de mener une expérimentation complémentaire plus approfondie de la librairie PANAME.

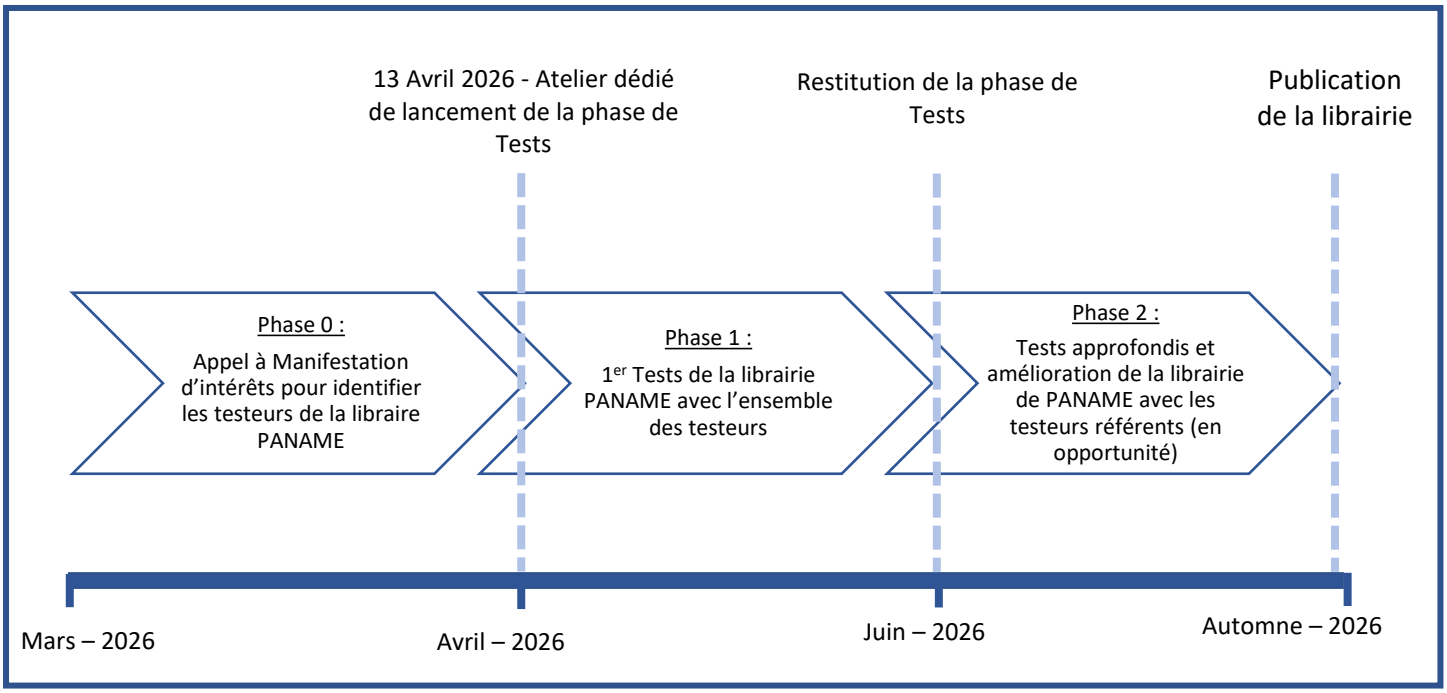

Figure 1 : cadrage et rétroplanning prévisionnel de la phase de test de la librairie PANAME

2.2. Objectifs et intérêts de l'AMI « Tests de la librairie PANAME »

Le présent AMI doit permettre d'identifier les acteurs (institutionnels et industriels) intéressés pour tester la librairie PANAME. Il s'adresse à tout entité publique et privée établie dans les États Membres de l'UE. Le consortium PANAME se réserve le droit de refuser certaines candidatures à partir d'un certain seuil ou bien si celles-ci sont jugées incomplètes et/ou hors-sujet.

La librairie PANAME est développée afin de donner un cadre adapté aux besoins des acteurs de l'écosystème pour implémenter leurs tests d'extraction de données d'entraînement, et permettre à ces derniers de conduire certaines évaluations nécessaires à l'analyse du statut au regard du RGPD des modèles qu'ils utilisent.
Cet AMI offre ainsi aux acteurs de l'écosystème une opportunité unique de tester en avant-première la librairie PANAME et de contribuer activement à son amélioration. En effet, les retours d'expérience permettront d'enrichir les fonctionnalités et d'assurer son adéquation aux besoins opérationnels.

Pour les participants, cette phase de tests permettra d'évaluer la confidentialité de leurs modèles d'IA, d'identifier d'éventuelles vulnérabilités et de renforcer la conformité de leurs solutions avec les exigences réglementaires en matière de protection des données personnelles.

La première phase de test sera ouverte à l'ensemble de l'écosystème ayant répondu à cet AMI suite à la présentation de la librairie lors d'un atelier dédié, sous réserve d'une réponse complète et jugée pertinente.

La 2nd phase de test sera restreint à un certain nombre d'acteurs sélectionné par la CNIL, le PEReN, l'ANSSI et Inria suite aux premiers tests effectués

2.3. Planning prévisionnel de l'AMI « Tests de la librairie PANAME »

- **10 Mars 2026 :** ouverture de l'AMI
- **9 Avril 2026 :** clôture de l'AMI
- **Du 26 Février 2026 jusqu'au 28 Mars 2026 :** instruction des dossiers déposés
- **13 avril 2026 après midi :** atelier en hybride à la CNIL pour les acteurs ayant répondu à l'AMI.

Ce planning prévisionnel doit permettre de lancer concrètement les premiers tests sur la librairie PANAME début avril 2026.

2.4. Nature des réponses attendues à l'AMI « Tests de la librairie PANAME »

Les réponses devront être complétées selon le format dédié « formulaire de réponse à l'AMI » disponible en annexe 1 du présent AMI en précisant notamment :

- **Les informations générales sur son organisation**
- **Les cas d'usages sur lesquels se positionne le candidat : c'est-à-dire les typologies des tests que le candidat sera capable d'effectuer avec la librairie PANAME en précisant notamment les données traités, les modèles d'IA utilisés et les scénarios d'attaques considérés.**

Les candidats au présent AMI sont également invités à préciser les éventuels prérequis et questions quant à l'utilisation de la librairie PANAME dans leur réponse à l'AMI.

2.5. Dépôt des candidatures à l'AMI « Tests de la librairie PANAME »

Les réponses à l'AMI devront être adressés en français, sous forme électronique, avant le 9 Avril 2026 (midi heure de Paris), à l'adresse mail : paname-test@inria.fr

3. Atelier organisé par les porteurs du projet PANAME suite au dépôt des candidatures

Suite à l'analyse des réponses à cet AMI, un atelier sera organisé par le consortium PANAME le 13 Avril 2026 à 14h en format hybride (une invitation sera transmise), avec l'intention de couvrir plusieurs objectifs :

1) Présenter la librairie PANAME et les types de tests attendus aux acteurs ayant répondu à l'AMI.
2) Faciliter la prise en main de la libraire PANAME aux acteurs ayant répondu à l'AMI et répondre à leurs questions.
3) Lancer officiellement la phase de tests de la libraire PANAME et partager les étapes de calendrier.

Cet atelier permettra également aux futurs testeurs de poser l'ensemble de leurs questions aux porteurs du projet PANAME concernant le projet et le fonctionnement de la librairie.

4. Respect du RGPD et cadre légal des tests

Les tests de la librairie PANAME qui seront réalisés dans le cadre du présent AMI (extraction et/ou réidentification de données sur les modèles d'IA) peuvent constituer, lorsqu'ils portent sur des modèles d'IA susceptibles d'avoir été entraînés sur des données à caractère personnel, un traitement de données au sens du Règlement général sur la protection des données (RGPD).

Dans le cadre du présent AMI, **les tests de la librairie PANAME peuvent être considérés comme poursuivant une finalité de recherche scientifique**. Le RGPD définit en effet la recherche scientifique de manière assez large, incluant par exemple le développement et la démonstration de technologies, la recherche fondamentale, la recherche appliquée… etc.
En l'espèce, ces tests d'exfiltration du modèle, d'analyse des sorties et, le cas échéant, de manipulation contrôlée de données extraites s'inscrivent dans une démarche de recherche scientifique puisqu'ils ont pour objet d'identifier ou d'évaluer, selon une méthodologie commune

- la capacité des modèles d'IA à compromettre des données personnelles par extraction ou réidentification.
- la mémorisation du modèle ;
- d'éventuelles vulnérabilités ;
- le statut juridique du modèle au regard du RGPD ;
- le niveau de conformité et la sécurité des solutions des participants.

Ces opérations s'inscrivent dans une démarche de production de connaissances structurées relatives aux vulnérabilités des modèles en matière de confidentialité. Les tests sont conduits dans un cadre méthodologique formalisé, reposant sur une grille d'analyse commune et reproductible.
Ils sont portés par un consortium d'acteurs publics (La CNIL, l'ANSSI, le PEReN et le projet IPoP du PEPR (programmes et équipements prioritaires de recherche) Cybersécurité piloté par Inria) à vocation scientifique et institutionnelle et ont pour objectif de produire des résultats transférables, notamment par la mise à disposition en source ouverte de la librairie PANAME et la diffusion des enseignements tirés des tests.

Chaque participant réalise les tests sur des modèles et des données qui auront été obtenues de manière licite**, en tant que responsable des traitement mis en œuvre dans ce cadre et de leur documentation.**
Les traitements de données personnelles qui seraient réalisés dans le cadre des tests de la librairie PANAME à des fins de recherche scientifique sont présumés compatibles avec la finalité initiale du traitement (entraînement du modèle d'IA). Chaque participant devra toutefois s'assurer:

- qu'aucune décision individuelle ne soit fondée sur les résultats des tests ;
- que des garanties appropriées pour les droits et libertés des personnes concernées soient mises en œuvre conformément à l'article 89 du RGPD (limiter les tests à la finalité de recherche poursuivie, réaliser les opérations dans un environnement sécurisé et contrôlé, restreindre les accès aux seules personnes habilitées, et assurer la traçabilité des opérations, etc.);
- que la base légale applicable soit identifiée et documentée.

5. Contact et information

Les questions concernant le présent AMI peuvent être adressées par courriel à l'adresse suivante : paname-test@inria.fr

Annexe 1 - Formulaire de réponse à l'AMI « Tests de la librairie PANAME »

Les projets sont à présenter sous la forme d'une **note de 3 à 4 pages par typologie de modèle** en respectant le formulaire ci-dessous. Si le candidat souhaite tester plusieurs modèles, il est invité à remplir plusieurs fois le formulaire en spécifiant chaque cas d'usage si nécessaire.

1. Informations générales

- **Nom de l'organisation** :
- **Personne référente** (nom, prénom, fonction, email, téléphone) :
- **Description succincte de l'organisation** (secteur d'activité, taille, expertise en IA) :
- **Motivation(s) pour tester cette librairie** (préciser succinctement ce qui vous amène à répondre à cet AMI) :

2. Description de la typologie des tests envisagés

1. **Types techniques de données traitées**

- **Modalité** (texte, image, audio, vidéo, données tabulaires, etc.) :
- **Format** (fichiers bruts, bases de données, API, etc.) :
- **Volume estimé** (nombre d'échantillons, taille en Go/To, etc.) :

2. **Catégorie de données personnelles ou confidentielles traitées**

- **Données personnelles** (identifiants, données biométriques, données de santé, etc.) :
- **Données sensibles ou confidentielles** (propriété intellectuelle, secrets industriels, etc.) :
- **Base légale du traitement d'entraînement du modèle** (consentement, intérêt légitime, obligation légale, recherche scientifique etc.) :

3. **Type de modèle entraîné**

- **Architecture** (réseaux de neurones, modèles de langage, forêts aléatoires, etc.) :
- **Taille du modèle** (nombre de paramètres, complexité, etc.) :
- **Méthode d'entraînement** (supervisé, non supervisé, auto-supervisé, etc.) :

4. **Finalité du modèle**

- **Objectif principal** (prédiction, classification, génération, etc.) :
- **Secteur d'application** (santé, finance, transport, etc.) :
- **Public cible** (grand public, professionnels, interne, etc.) :

5. **Modélisation du risque d'extraction de données du modèle (optionnel)**

   Avez-vous déjà conduit une modélisation des risques d'extraction de données du modèle ? Si oui :

- **Méthodologie utilisée** pour évaluer les risques (analyse théorique, tests empiriques, etc.) :
- **Résultats préliminaires** (vulnérabilités identifiées, niveau de risque estimé) :

6. **Types de tests d'extraction envisagés**

- **Scénarios d'extraction** (*Attribute inference attack, model inversion -reconstruction-, membership inference attack,* etc.) :

7. **Modalité d'accès au modèle cible**

   Dans le cadre de l'utilisation de la librairie Paname, quels formats vous paraissent les plus adaptés pour accéder à votre modèle en inférence (sorties du modèle, accès à des valeurs internes comme les logits, la valeur de la fonction de coût, ou encore les gradients de la dernière couche) ? Merci de préciser toutes options techniques possibles parmi :

   - **Chargement direct** du modèle en mémoire
   - Appel via une **API interne**

- Accès à des **prédictions précalculées**
- **Autre** : préciser.
- Quelles sont les contraintes techniques supplémentaires que vous anticipez ?

8. Les modèles proxy ou *shadow models* (optionnel)

La conduite de certaines attaques demande de disposer d'un ou plusieurs modèles proxy, dits *shadow models*. Serez-vous en mesure de fournir de tels models proxy en parallèle du modèle cible ?

- Si **non**, quels serait le principal facteur limitant ? (ressources de calcul, infrastructure de code, framework logiciel)
- Si **oui**, sous quelle forme mettriez-vous ces modèles à disposition de la librairie ? (Chargement direct du modèle en mémoire, API, accès à des prédictions précalculées, autre: préciser)
- Quelles sont les **contraintes techniques supplémentaires** que vous anticipez ?

9. Ressources de calcul et temps homme

- **Type d'infrastructure** (cloud, on-premise, GPU/TPU, etc.) :
- **Capacité disponible** (nombre de machines, puissance de calcul, etc.) :
- **Équipe dédiée** (nombre de personnes, compétences) :
- **Temps estimé** pour la participation au test (en jours/hommes) :

3. Préciser les éventuels prérequis et questions quant à l'utilisation de la librairie PANAME

Annexe 2 - Introduction aux techniques d'extraction de données d'entraînement

1. La persistance de la donnée dans le modèle

Lors de sa phase d'entraînement, un modèle d'apprentissage automatique, et a fortiori un réseau de neurones profond (Deep Learning), ne se contente pas d'apprendre des règles statistiques. Il agit comme un système de compression d'informations qui peut conserver, encoder et mémoriser des traces granulaires des données individuelles lui ayant été soumises.

Un tiers n'ayant aucun accès à la base de données d'entraînement, mais ayant un accès total ou partiel au modèle (via un accès aux poids, une API publique ou une application), peut exploiter ce point d'entrée pour extraire des informations à propos de la base de donnée d'entraînement.

2. Généralisation et mémorisation

Pour comprendre intuitivement comment les techniques d'extraction peuvent fonctionner, on peut aborder le problème sous l'angle de l'équilibre entre généralisation et mémorisation.

- La **Généralisation** : Le modèle apprend les caractéristiques communes d'une classe d'objets (ex: la structure générale d'un visage humain) pour reconnaître de nouveaux visages qu'il n'a jamais vus.

- La **Mémorisation** : Lorsque le modèle est trop complexe ou entraîné trop longtemps sur un jeu de données limité, il commence à apprendre "par cœur" des exemples spécifiques. Il ne retient pas seulement "ce qu'est un visage", mais "le visage précis de l'individu n°4589".

Un modèle se comporte mathématiquement différemment face à une donnée qu'il a observée pendant l'entraînement (et possiblement "apprise par cœur") par rapport à une donnée inconnue. Par exemple, il peut traiter la donnée mémorisée avec une confiance extrême et une erreur quasi-nulle. Cette différence de comportement (ce "signal") peut alors être exploité par un attaquant..

3. Les grands types de techniques d'extraction

Nous détaillons ici le fonctionnement technique des trois principales familles d'attaques.

Attaques par Inférence d'Appartenance (*Membership Inference Attack* - MIA)

C'est l'attaque la plus fondamentale. Elle vise à répondre à une simple question binaire : « Les données concernant l'individu X ont-elles été utilisées pour entraîner ce modèle ? »
Différentes stratégies peuvent-être employées, comme par exemple:

- L'observation des scores de confiance (*Confidence Scores*) :

La plupart des modèles de classification (ex: reconnaissance d'images, diagnostic médical) ne répondent pas par un simple "Oui/Non". Ils fournissent un vecteur de probabilités (ex: "Malade : 99.8%", "Sain : 0.2%").
Si l'attaquant soumet le dossier de Monsieur X au modèle et que celui-ci renvoie un score de confiance anormalement élevé (proche de 100%), c'est un indice fort que le modèle a déjà vu cette donnée exacte durant son entraînement. Face à une donnée inconnue, le modèle hésite davantage.

- L'utilisation de "Modèles Proxy" (*Shadow Models*) :

Pour confirmer cette intuition, l'attaquant utilise une technique avancée. Il crée plusieurs de ses propres modèles (les modèles proxy) qu'il entraîne sur des données qu'il contrôle (certaines qu'il inclut, d'autres qu'il exclut).
Il observe comment ses modèles réagissent face aux données incluses (réponse A) et aux données exclues (réponse B).
Il entraîne ensuite un classifieur d'attaque : une IA dont le seul but est de reconnaître si le comportement du modèle cible ressemble au comportement "A" (donnée membre) ou "B" (donnée non-membre).

Si l'appartenance à la base de données révèle une information sensible (ex: base de données d'une application de rencontres, registre de patients psychiatriques), la vie privée de l'individu peut-être compromise.

Attaque par Inférence d'Attribut (*Attribute Inference Attack*)

Ici, l'attaquant possède une connaissance partielle des données individuelles utilisées pour l'entraînement et cherche à inférer des attributs supplémentaires sur de ces données en utilisant le modèle comme un oracle.

L'attaque repose sur le fait que le modèle encode des corrélations fines entre différents attributs des données d'entraînement. Ces attaques fonctionnent de la manière suivante :

- Préparation de l'entrée : L'attaquant dispose d'un profil incomplet de la victime (ex: Âge, Sexe, Code Postal) mais il lui manque l'attribut sensible (ex: Revenu Fiscal).

- Interrogation itérative : L'attaquant peut tester toutes les valeurs possibles de l'attribut manquant en interrogeant le modèle.

- Essai 1 : [Âge, Sexe, CP, Revenu = 20k€] -> Le modèle donne une réponse cohérente mais avec une confiance moyenne.

- Essai 2 : [Âge, Sexe, CP, Revenu = 80k€] -> Le modèle réagit avec une confiance très élevée.

Contrairement à une simple déduction statistique (qui dirait "les gens de ce quartier sont souvent riches"), l'inférence d'attribut via un modèle d'IA vise à inférer des informations individuelles sur les données présentes dans l'ensemble d'entraînement.

Attaques par Reconstruction / Inversion de Modèle (Model Inversion Attack)

Dans la situation où le modèle cible n'est pas génératif (classification, régression, etc.) il s'agit du type de technique le plus sophistiqué. L'objectif est de régénérer une approximation de la donnée brute (le visage, l'empreinte digitale, le texte) à partir d'un accès au modèle, sans jamais avoir vu l'entrée originale.

Exemple de fonctionnement technique détaillé : L'optimisation par remontée de gradient

L'attaquant ne "cherche" pas l'image dans le modèle (elle n'y est pas stockée comme dans un disque dur). Il la reconstruit progressivement, en l'approchant de manière itérative à partir des informations encodées dans le modèle.

- Initialisation :
  L'attaquant commence avec une image composée de "bruit" (des pixels gris aléatoires) ou un texte vide.

- Inférence :
  - L'attaquant soumet ce bruit au modèle en demandant : "Quelle est la probabilité que ceci soit Monsieur Dupont ?".
  - Le modèle répond : "0.01%".
  - Mais le modèle fournit (ou l'attaquant peut estimer) ce qu'on appelle un gradient. Le gradient est une direction mathématique qui indique : "Pour augmenter ce score, il faudrait éclaircir ces pixels-là et assombrir ceux-ci".

- L'itération (remontée de gradient) :
  - L'attaquant modifie légèrement l'image de bruit en suivant les indications du gradient.
  - Il soumet la nouvelle image. Le modèle répond : "0.05%".

Petit à petit, l'image de bruit converge vers une forme qui maximise la reconnaissance du modèle pour "Monsieur Dupont". Le résultat final est une reconstruction visuelle (souvent floue mais identifiable) du visage qui a servi à l'entraînement.