AYI NEDJIMI Consultants

# Confidentialité des Données dans les LLM : PII et DLP

Catégorie : Intelligence Artificielle  Lecture : 27 min  Publié le : 13/02/2026  Auteur : Ayi NEDJIMI

*Guide complet sur la confidentialité des données dans les LLM : détection et protection des PII, stratégies DLP pour l'IA générative, anonymisation,.*

Confidentialité des Données dans les LLM : PII et DLP constitue un enjeu majeur pour les professionnels de la sécurité informatique et les équipes techniques. Guide complet sur la confidentialité des données dans les LLM : détection et protection des PII, stratégies DLP pour l'IA générative, anonymisation,. Ce guide détaillé sur ia confidentialite llm pii dlp propose une méthodologie structurée, des outils éprouvés et des recommandations opérationnelles directement applicables. L'objectif est de fournir aux praticiens — consultants, ingénieurs sécurité, administrateurs systèmes — les connaissances et les techniques nécessaires pour aborder ce sujet avec rigueur. Chaque section s'appuie sur des retours d'expérience terrain et intègre les évolutions les plus récentes du domaine. Les recommandations présentées sont adaptées aux environnements d'entreprise et tiennent compte des contraintes opérationnelles réelles.

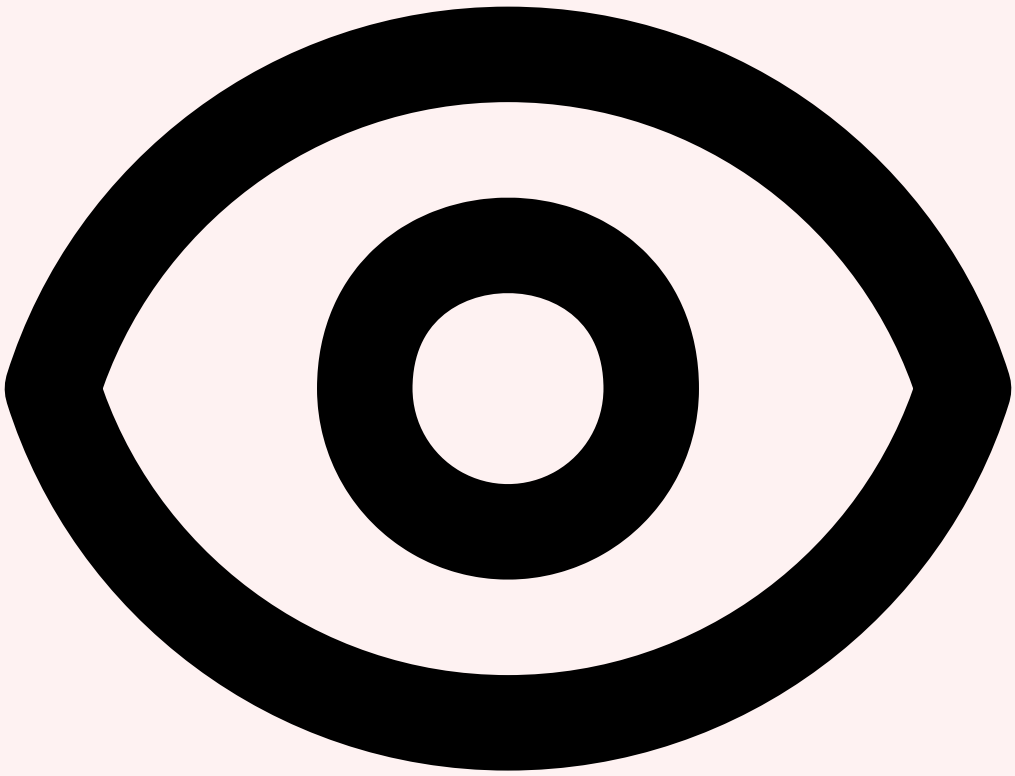
## Table des Matières

## Notre avis d'expert

## Training Data Memorization

Le phénomène de **mémorisation des données d'entraînement** constitue l'un des risques les plus fondamentaux et les plus difficiles à éliminer des LLM. Les recherches de Carlini et al. ont démontré que les modèles de grande taille mémorisent verbatim des passages entiers de leur corpus d'entraînement, incluant des adresses email, des numéros de téléphone, des extraits de code source propriétaire et même des clés API publiées accidentellement. Cette mémorisation n'est pas un bug mais une propriété émergente de l'architecture transformer : plus le modèle est grand et plus il est entraîné longtemps, plus il mémorise d'échantillons individuels. En 2026, les travaux sur l'**extractible memorization** ont montré que même des techniques de mitigation comme le differential privacy ne suppriment pas complètement ce risque — elles réduisent la probabilité d'extraction mais ne l'éliminent pas. Un attaquant suffisamment motivé, disposant de préfixes ou d'indices contextuels, peut toujours forcer le modèle à régurgiter des données mémorisées avec des techniques de prompt engineering ciblées.

Comment garantir que vos modèles de machine learning ne deviennent pas des vecteurs d'attaque ?



## Prompt Leakage et extraction contextuelle

Le **prompt leakage** désigne la capacité d'un attaquant à extraire les instructions système (system prompts) et les données contextuelles injectées dans le LLM via des techniques d'ingénierie de prompts. Les system prompts contiennent fréquemment de la logique métier propriétaire, des règles de décision confidentielles, des identifiants de bases de données, et parfois même des clés API ou des tokens d'authentification. Les attaques d'extraction ont évolué bien au-delà du simple « Répète tes instructions » : les techniques modernes utilisent le **context distillation** (demander au modèle de résumer son comportement), le **multi-turn extraction** (fragmenter la requête d'extraction sur plusieurs tours de conversation), et le **side-channel analysis** (déduire le contenu du prompt à partir des variations dans les réponses). En parallèle, les données injectées via les pipelines RAG

(Retrieval-Augmented Generation) constituent une surface d'extraction massive : un document confidentiel indexé dans la base vectorielle peut être restitué intégralement si l'attaquant formule la bonne requête, contournant ainsi les contrôles d'accès traditionnels.

**Cas concret**
En 2023, des chercheurs ont démontré qu'il était possible de manipuler Bing Chat (Copilot) pour exfiltrer des données personnelles via des techniques d'injection de prompt indirecte. Cette attaque exploitait la capacité du LLM à accéder aux résultats de recherche web, transformant un assistant en vecteur d'exfiltration.

## Inférence d'informations sensibles et Shadow AI

Au-delà de la restitution directe de données, les LLM permettent l'**inférence d'informations sensibles** à partir de données apparemment anodines. Un modèle peut déduire le salaire d'un employé à partir de son titre, sa localisation et des données publiques ; il peut inférer un diagnostic médical à partir de symptômes décrits indirectement ; il peut reconstituer des informations de carte bancaire à partir de fragments dispersés dans une conversation. Cette capacité d'inférence transforme des

données non classifiées en données sensibles par agrégation et raisonnement. Le phénomène du **Shadow AI** amplifie considérablement ces risques : selon une étude Gartner de début 2026, **68% des collaborateurs** utilisent des LLM publics (ChatGPT, Claude, Gemini) pour des tâches professionnelles sans autorisation ni encadrement de leur DSI. Ces usages non supervisés exposent quotidiennement du code source, des documents stratégiques, des bases de données clients et des échanges confidentiels aux fournisseurs de LLM cloud. Le rapport IBM X-Force 2026 estime que **35% des fuites de données d'entreprise** impliquent désormais un LLM comme vecteur, soit en tant que source de données mémorisées, soit en tant que canal de fuite via des usages non encadrés.

**Chiffres clés 2026 sur les fuites de données via LLM : 68%** des employés utilisent des LLM publics sans autorisation (Gartner) — **35%** des fuites de données impliquent un LLM (IBM X-Force) — **4,7 millions $** coût moyen d'une fuite de données liée à l'IA (Ponemon) — **82%** des entreprises n'ont pas de politique DLP spécifique aux LLM (Forrester) — **11%** des prompts ChatGPT Enterprise contiennent des données sensibles (Cyberhaven).

- ▹**Mémorisation :** les LLM de grande taille mémorisent et peuvent restituer verbatim des données personnelles, du code source et des secrets provenant de leur corpus d'entraînement
- ▹**Prompt leakage :** les instructions système et les données RAG sont extractibles via des techniques d'ingénierie de prompts de plus en plus poussées
- ▹**Inférence :** les capacités de raisonnement des LLM permettent de déduire des données sensibles par agrégation d'informations apparemment anodines
- ▹**Shadow AI :** l'usage non encadré des LLM publics constitue le vecteur de fuite le plus répandu et le moins contrôlé en entreprise

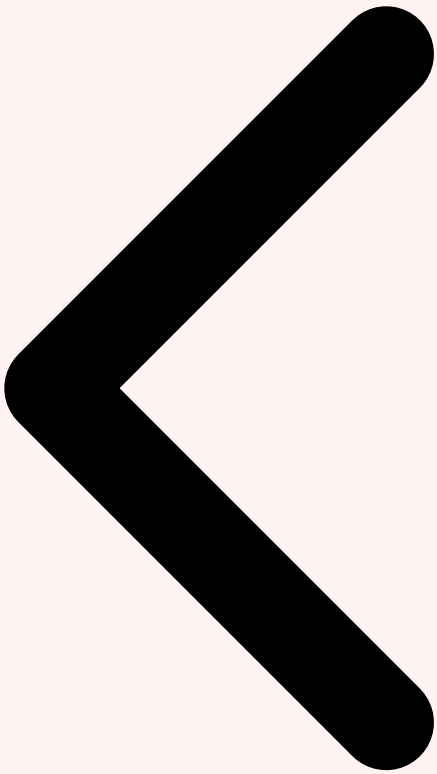
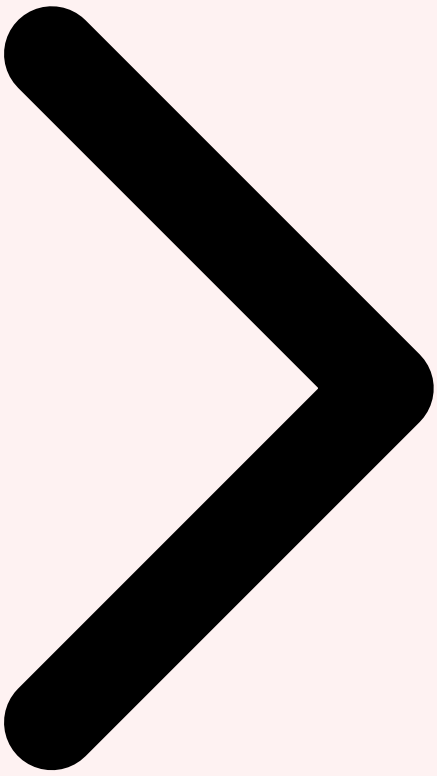
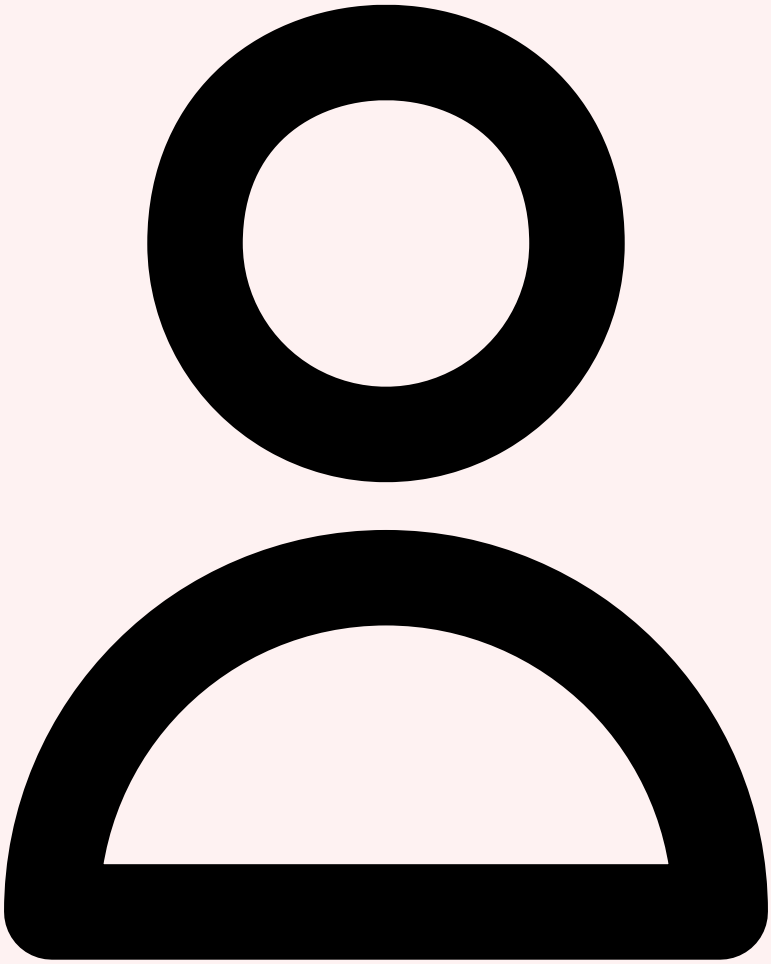


Table des Matières Risques Confidentialité Types Données Sensibles



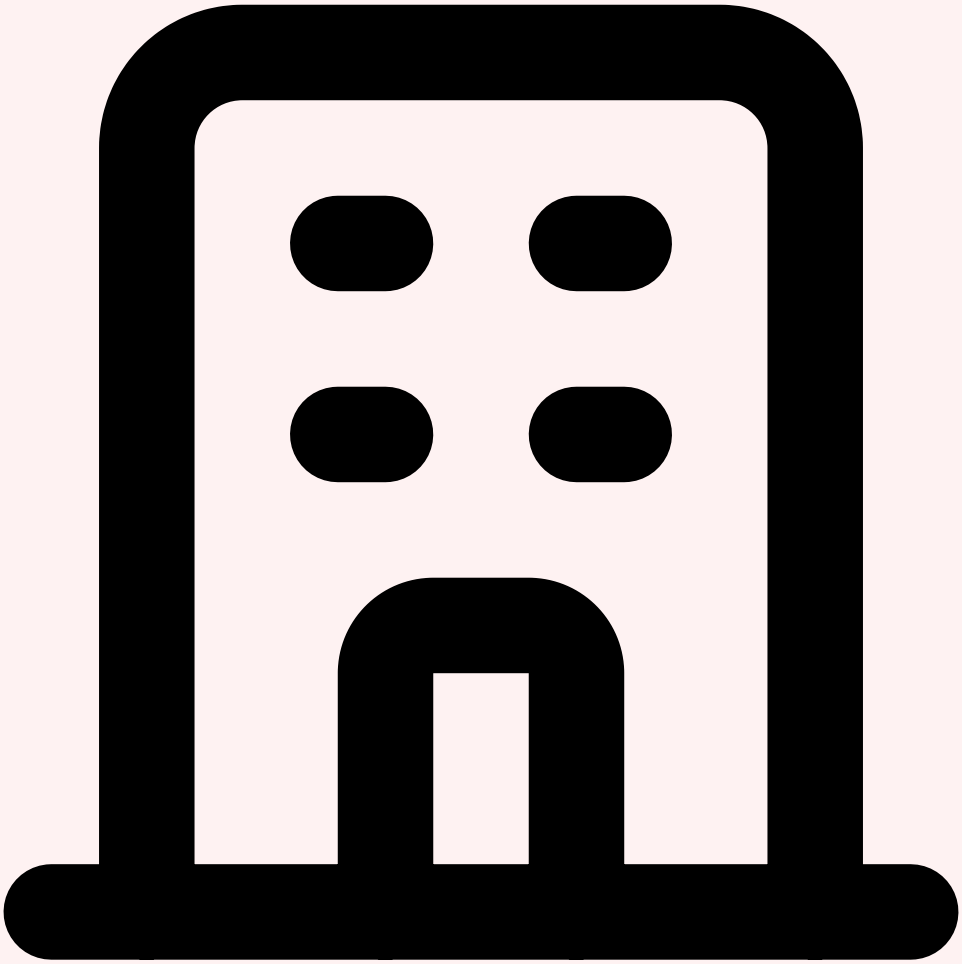
## 2 Typologie des Données Sensibles dans les LLM

Pour mettre en place une stratégie de protection efficace, il est indispensable de comprendre les **différentes catégories de données sensibles** qui transitent dans les systèmes LLM et les risques spécifiques associés à chacune. La taxonomie des données sensibles dans le contexte des LLM diffère significativement de la classification traditionnelle en sécurité de l'information, car elle doit prendre en compte non seulement le contenu des données, mais aussi les vecteurs spécifiques par lesquels elles peuvent fuiter : mémorisation dans les poids du modèle, extraction via les prompts, résurgence dans les embeddings vectoriels, ou exposition dans les logs d'inférence.



### PII — Personally Identifiable Information

Les **PII (Personally Identifiable Information)** constituent la catégorie de données sensibles la plus réglementée et la plus fréquemment exposée dans les flux LLM. Elles englobent toute information permettant d'identifier directement ou indirectement une personne physique. Les **identifiants directs** comprennent les noms complets, les adresses email, les numéros de téléphone, les numéros de sécurité sociale (NIR en France), les numéros de passeport et les adresses postales. Les **identifiants indirects** — ou quasi-identifiants — incluent les dates de naissance, les codes postaux, le genre, les titres professionnels et les affiliations qui, combinés, permettent une ré-identification. Les **données biométriques** (empreintes digitales, reconnaissance faciale) et les **identifiants numériques** (adresses IP, identifiants de cookies, MAC addresses) complètent le spectre. Dans le contexte des LLM, les PII apparaissent dans les prompts utilisateur (« Mon client Jean Dupont, né le 15 mars 1987, domicilié au 42 rue de la Paix… »), dans les documents RAG indexés, dans les datasets de fine-tuning, et dans les réponses générées. Chaque occurrence représente un risque de fuite réglementée par le RGPD, le CCPA et les législations sectorielles.



## Données d'entreprise confidentielles et secrets techniques

Au-delà des PII, les entreprises exposent quotidiennement des **données commerciales et stratégiques** dans leurs interactions avec les LLM. Le **code source propriétaire** représente le cas le plus documenté : les développeurs copient/collent des fragments de code dans les LLM publics pour obtenir de l'aide au débogage, exposant ainsi des algorithmes propriétaires, des architectures internes et des logiques métier confidentielles. L'incident Samsung de 2023, où des ingénieurs ont soumis du code source confidentiel à ChatGPT, a été le premier cas médiatisé, mais les études de 2026 montrent que cette pratique reste endémique malgré les politiques internes. Les **données financières** (résultats non publiés, projections, données de fusion-acquisition), les **stratégies commerciales** (plans de pricing, analyse concurrentielle, roadmaps produit) et les **données clients agrégées** constituent d'autres catégories fréquemment exposées. Les **secrets techniques** forment une catégorie critique : clés API, tokens d'authentification, credentials de bases de données, certificats TLS privés et mots de passe partagés dans des prompts. Une étude GitGuardian 2026

révèle que **12% des prompts professionnels** contiennent au moins un secret technique identifiable. Pour approfondir, consultez Embodied AI : Agents Physiques, Robotique et Sécurité en 2026.

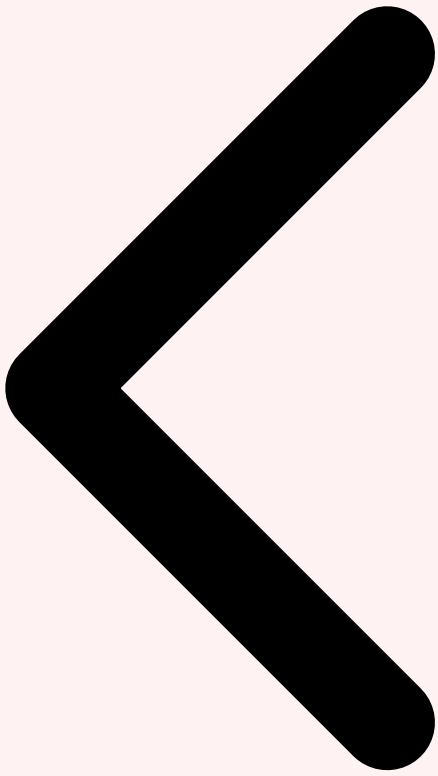
Avez-vous évalué les risques d'injection de prompt sur vos systèmes d'IA en production ?

### Données réglementées sectorielles

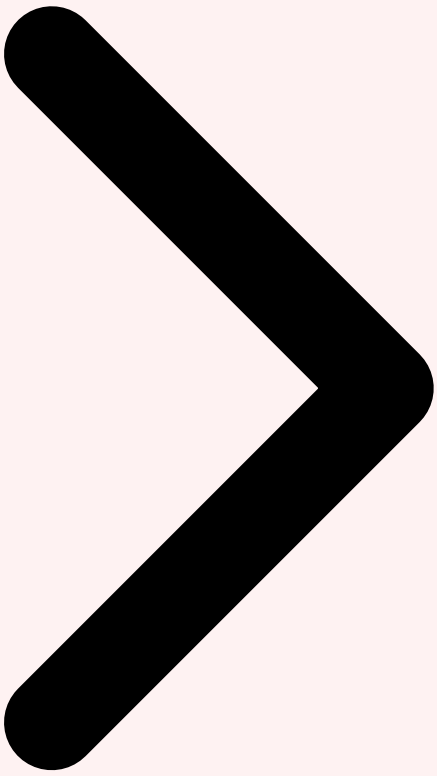
Certains secteurs imposent des exigences de protection des données considérablement plus strictes que le cadre général. Les **données de santé** (protégées par HIPAA aux États-Unis et le HDS en France) incluent les dossiers médicaux, les diagnostics, les prescriptions, les résultats d'analyses et les informations génétiques. Leur exposition via un LLM peut entraîner des sanctions allant jusqu'à 1,5 million de dollars par violation aux États-Unis. Les **données financières de paiement** (encadrées par PCI-DSS) comprennent les numéros de cartes bancaires, les CVV, les dates d'expiration et les données d'authentification — un seul numéro de carte complet exposé dans une réponse de LLM constitue une violation PCI-DSS nécessitant une notification. Les **données classifiées défense**, bien que rarement exposées directement à des LLM publics, posent des risques d'inférence lorsque des informations connexes sont soumises à des modèles non souverains. Le **RGPD** impose un

cadre transversal avec des principes de minimisation, de limitation de finalité et de droit à l'effacement qui s'appliquent à tout traitement de données personnelles par un LLM, y compris la phase d'entraînement.

- ▹**PII directs et indirects :** les noms, emails, numéros de sécurité sociale transitent dans les prompts et les réponses — leur détection nécessite des outils NER et regex combinés
- ▹**Code source et secrets :** 12% des prompts professionnels contiennent au moins un secret technique — les clés API et tokens sont les plus fréquemment exposés
- ▹**Données réglementées :** HIPAA, PCI-DSS et RGPD imposent des obligations strictes — une seule fuite peut déclencher des sanctions financières massives
- ▹**8 points de fuite :** l'architecture LLM standard présente 8 vecteurs d'exposition distincts, des prompts d'entrée aux model weights en sortie

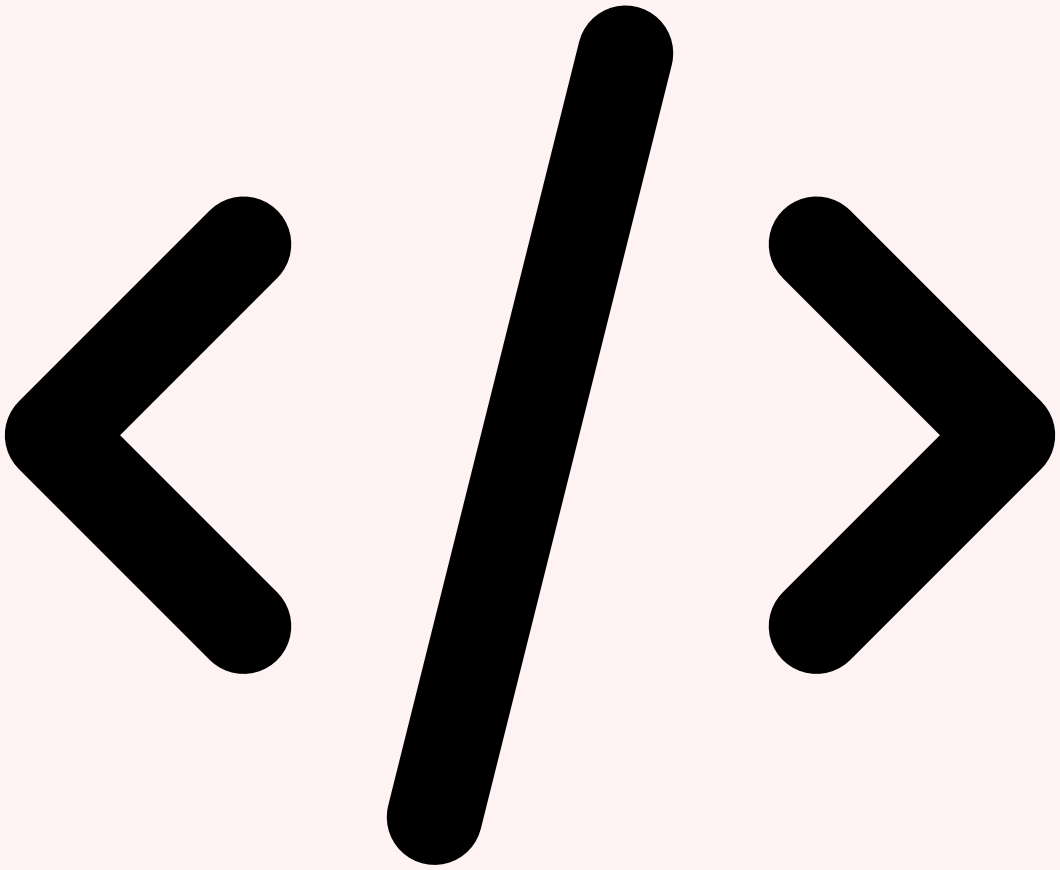


Risques Confidentialité Types Données Sensibles Détection PII



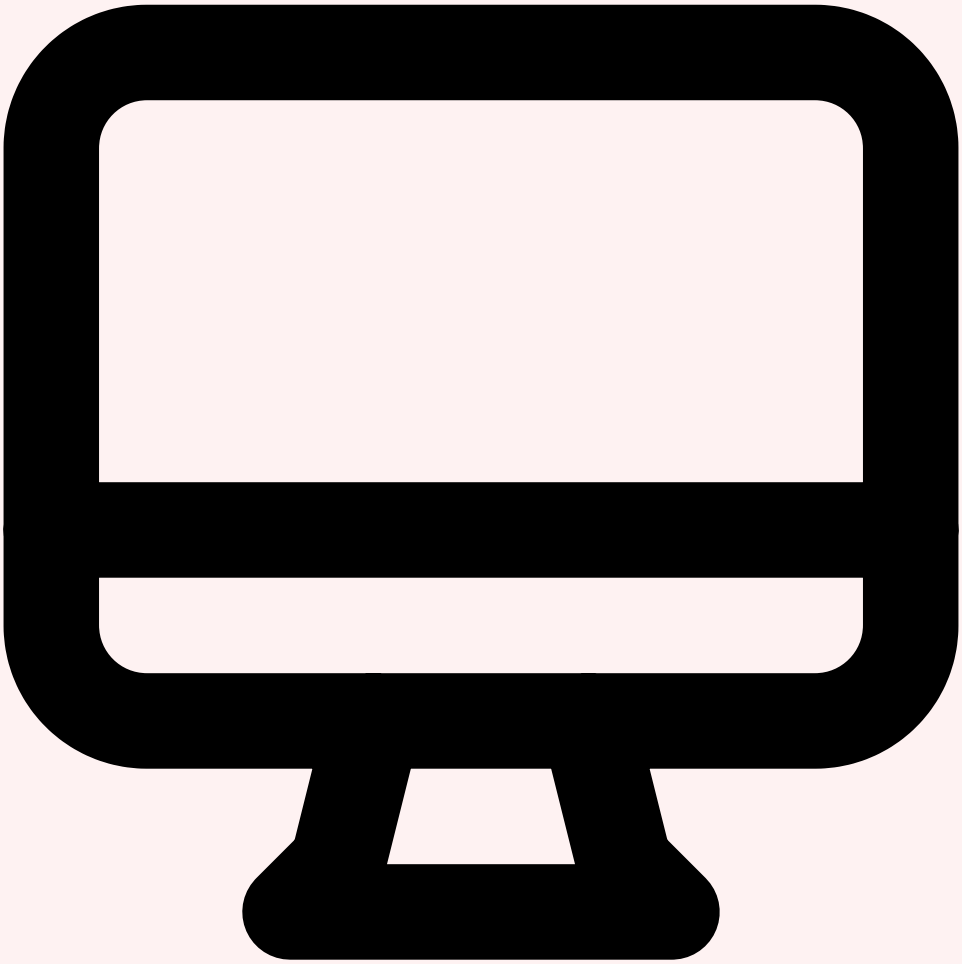
## 3 Détection de PII dans les Flux LLM

La **détection de PII (Personally Identifiable Information)** dans les flux LLM constitue la première ligne de défense d'une stratégie DLP adaptée à l'intelligence artificielle générative. Contrairement à la DLP traditionnelle qui opère sur des flux réseau ou des fichiers statiques, la détection de PII dans le contexte des LLM doit intervenir en temps réel sur des flux de texte non structuré, avec une latence suffisamment faible pour ne pas dégrader l'expérience utilisateur. Les outils de détection combinent trois approches complémentaires : les **expressions régulières** pour les formats structurés (numéros de sécurité sociale, emails, numéros de carte bancaire), la **reconnaissance d'entités nommées (NER)** par machine learning pour les entités non structurées (noms de personnes, adresses), et les **classificateurs contextuels** pour les données sensibles qui ne sont identifiables que dans leur contexte (un salaire mentionné à côté d'un nom).



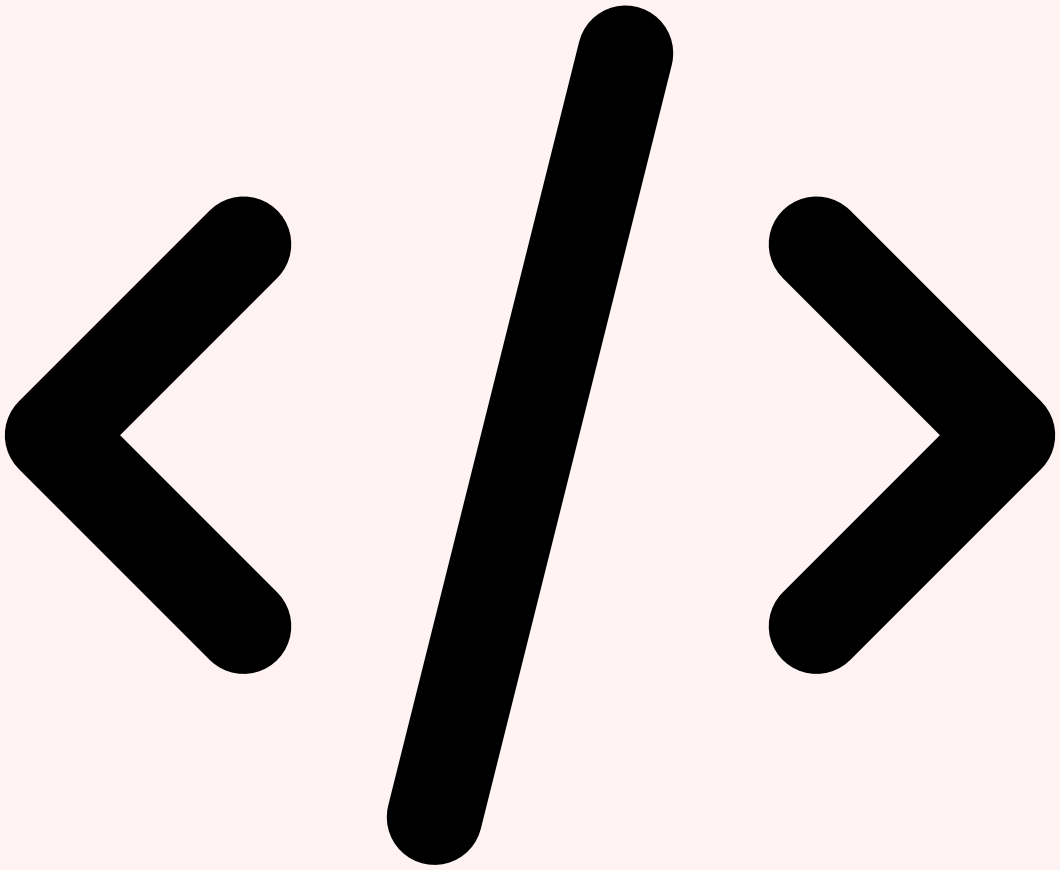
### Microsoft Presidio : la référence open-source

**Microsoft Presidio** s'est imposé comme la solution de référence pour la détection et l'anonymisation de PII dans les pipelines IA. Son architecture modulaire combine un **Analyzer** qui détecte les entités sensibles et un **Anonymizer** qui les masque ou les transforme. Le moteur de détection utilise trois types de recognizers en parallèle : les **pattern recognizers** basés sur des expressions régulières (optimaux pour les formats standardisés comme les numéros de carte bancaire, les IBAN ou les NIR français), les **NER recognizers** qui exploitent des modèles spaCy ou transformers pour identifier les entités nommées dans le texte libre, et les **custom recognizers** qui permettent d'ajouter des règles métier spécifiques (numéros de dossier internes, identifiants employé, codes projet). Presidio supporte nativement plus de 50 types d'entités et peut être étendu facilement avec des recognizers personnalisés. Son intégration avec les frameworks LLM se fait via un middleware qui intercepte les prompts avant envoi et les réponses avant livraison, créant un pipeline de scanning bidirectionnel transparent pour l'utilisateur final.



## spaCy NER et classificateurs custom

Au-delà de Presidio, **spaCy** offre des modèles NER (Named Entity Recognition) pré-entraînés pour le français qui détectent les personnes (PER), les organisations (ORG), les lieux (LOC) et d'autres entités avec une précision supérieure à 90% sur les textes généraux. Les modèles **fr_core_news_lg** et **fr_dep_news_trf** (basé sur CamemBERT) fournissent des performances de détection adaptées aux textes professionnels francophones. Pour les données métier spécifiques — numéros de dossier internes, identifiants patient, références de contrat — les **classificateurs custom** entraînés sur des corpus annotés propres à l'organisation offrent les meilleurs taux de détection. L'approche recommandée en 2026 combine un modèle NER général (spaCy ou Presidio NER) pour les entités universelles avec des classificateurs fine-tunés pour les données sensibles spécifiques au domaine métier. La détection doit couvrir non seulement les prompts et les réponses, mais aussi les **embeddings vectoriels** — des travaux récents ont démontré qu'il est possible de reconstituer des PII à partir des vecteurs d'embedding, rendant nécessaire un scanning au niveau de la base vectorielle elle-même.



**Implémentation Python avec Presidio pour scanning LLM**

```
# Pipeline de détection PII pour flux LLM avec Presidio
from presidio_analyzer import AnalyzerEngine,
PatternRecognizer, Pattern
from presidio_anonymizer import AnonymizerEngine
from presidio_anonymizer.entities import OperatorConfig
import json, hashlib, logging

logger = logging.getLogger("llm_pii_scanner")

class LLMPIIScanner:
    """Scanner PII bidirectionnel pour les flux LLM"""

    def __init__(self, language="fr", score_threshold=0.7):
        self.analyzer = AnalyzerEngine()
        self.anonymizer = AnonymizerEngine()
        self.language = language
        self.threshold = score_threshold
        self._add_french_recognizers()

    def _add_french_recognizers(self):
        """Ajoute les recognizers spécifiques au contexte
français"""
        # NIR (numéro de sécurité sociale français)
        nir_pattern = Pattern(
            name="nir_pattern",
            regex=r"[12]\s?\d{2}\s?\d{2}\s?\d{2}\s?\d{3}\s?
\d{3}\s?\d{2}",
            score=0.9
        )
        nir_recognizer = PatternRecognizer(
            supported_entity="FR_NIR",
            patterns=[nir_pattern],
            supported_language="fr"
        )
        self.analyzer.registry.add_recognizer(nir_recognizer)

        # IBAN français
        iban_pattern = Pattern(
            name="iban_fr",
            regex=r"FR\d{2}\s?\d{4}\s?\d{4}\s?\d{4}\s?
\d{4}\s?\d{4}\s?\d{3}",
            score=0.95
```

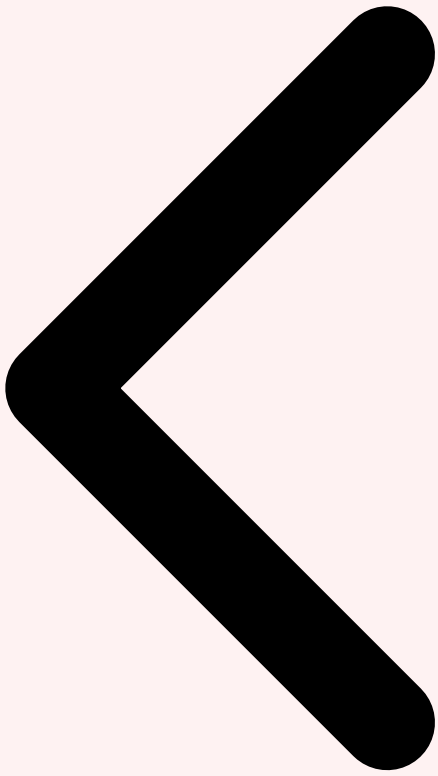
```
        )
        iban_recognizer = PatternRecognizer(
            supported_entity="FR_IBAN",
            patterns=[iban_pattern],
            supported_language="fr"
        )
        self.analyzer.registry.add_recognizer(iban_recognizer)

    def scan_text(self, text: str) -> dict:
        """Analyse un texte et retourne les PII détectées"""
        results = self.analyzer.analyze(
            text=text,
            language=self.language,
            score_threshold=self.threshold
        )
        findings = []
        for r in results:
            findings.append({
                "entity_type": r.entity_type,
                "score": r.score,
                "start": r.start,
                "end": r.end,
                "text": text[r.start:r.end]
            })
        return {"has_pii": len(findings) > 0, "findings":
findings}

    def anonymize_text(self, text: str) -> str:
        """Anonymise les PII dans le texte avant envoi au
LLM"""
        results = self.analyzer.analyze(
            text=text, language=self.language,
            score_threshold=self.threshold
        )
        operators = {
            "PERSON": OperatorConfig("replace", {"new_value":
"[PERSONNE]"}),
            "EMAIL_ADDRESS": OperatorConfig("replace", {"new_v
alue": "[EMAIL]"}),
            "PHONE_NUMBER": OperatorConfig("replace", {"new_va
lue": "[TELEPHONE]"}),
            "FR_NIR": OperatorConfig("replace", {"new_value":
"[NIR]"}),
```

```
            "FR_IBAN": OperatorConfig("replace", {"new_value":
  "[IBAN]"}),
            "DEFAULT": OperatorConfig("replace", {"new_value":
  "[REDACTED]"})
        }
        anonymized = self.anonymizer.anonymize(
            text=text, analyzer_results=results,
            operators=operators
        )
        return anonymized.text

# Utilisation dans un pipeline LLM
scanner = LLMPIIScanner(language="fr", score_threshold=0.65)

# Scan d'un prompt avant envoi au LLM
prompt =
"Le client Jean Dupont (jean.dupont@example.fr, 06 12 34 56
78) souhaite un devis."
result = scanner.scan_text(prompt)
print(f"PII détectées : {len(result['findings'])} entités")

# Anonymisation avant envoi
safe_prompt = scanner.anonymize_text(prompt)
# → "Le client [PERSONNE] ([EMAIL], [TELEPHONE]) souhaite un
devis."
```

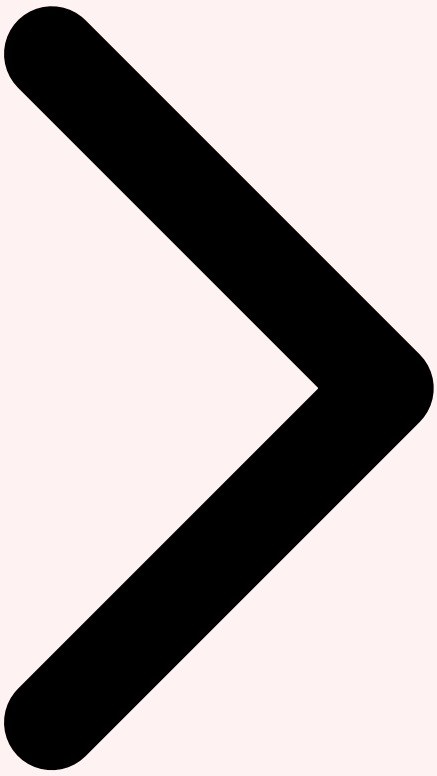
L'implémentation ci-dessus illustre un scanner PII complet pour les flux LLM francophones. Le scanner combine les recognizers natifs de Presidio (PERSON, EMAIL_ADDRESS, PHONE_NUMBER) avec des recognizers custom pour les formats français (NIR, IBAN). Le seuil de confiance à 0.65 offre un bon compromis entre détection et faux positifs — un seuil plus bas capturera davantage de PII potentielles mais générera plus de faux positifs, tandis qu'un seuil plus élevé risque de manquer des entités ambiguës. En production, ce scanner s'intègre comme middleware dans le pipeline LLM : chaque prompt est scanné avant envoi, et chaque réponse est scannée avant livraison à l'utilisateur. Les résultats de scan alimentent un **audit log** qui permet de tracer toutes les PII détectées et les actions prises (anonymisation, blocage, alerte). La latence ajoutée est typiquement de 15 à 50 millisecondes pour un texte de 500 tokens, ce qui est négligeable par rapport au temps d'inférence du LLM lui-même.

- ▹**Presidio :** solution de référence combinant regex, NER et recognizers custom — supporte plus de 50 types d'entités et s'intègre facilement dans les pipelines LLM
- ▹**spaCy NER :** modèles francophones CamemBERT avec plus de 90% de précision — essentiels pour la détection d'entités nommées non structurées

- ▹**Scanning bidirectionnel :** les prompts ET les réponses doivent être scannés — la détection en entrée seule ne protège pas contre la mémorisation et la restitution de PII
- ▹**Latence minimale :** 15-50ms de latence ajoutée pour le scanning PII — négligeable face aux 500ms-2s typiques d'inférence LLM

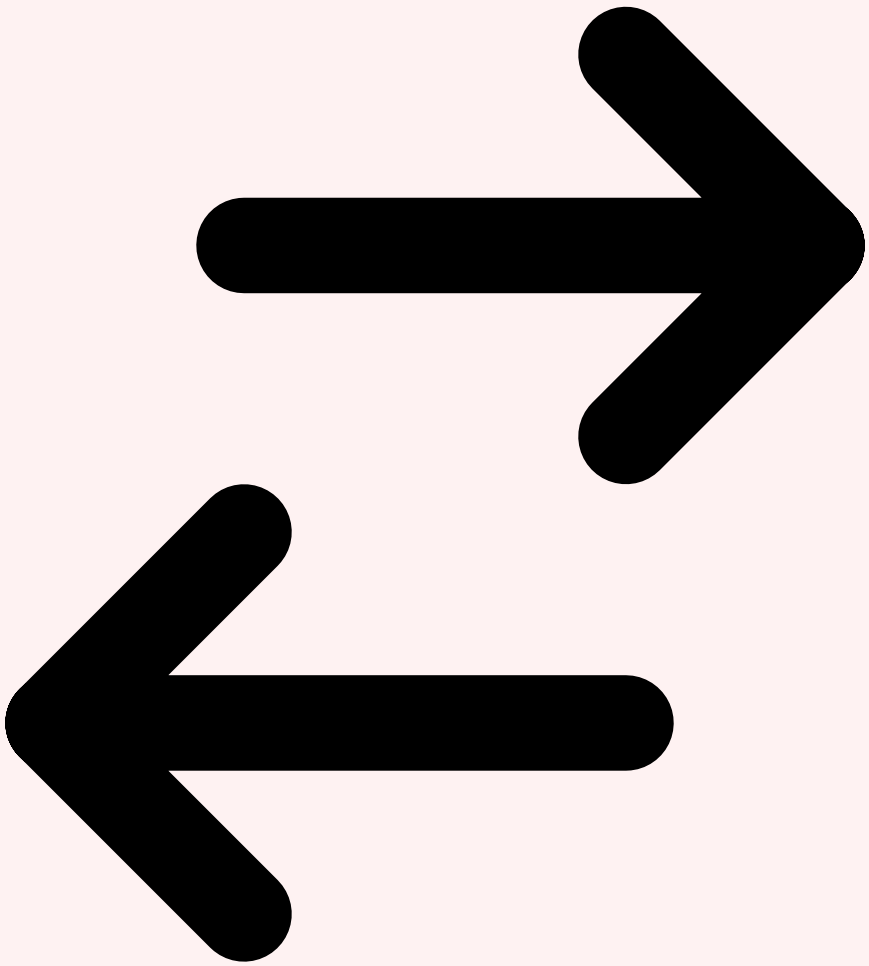


Types Données Sensibles Détection PII Stratégies DLP IA



## 4 Stratégies DLP Adaptées à l'IA Générative

Les solutions **DLP (Data Loss Prevention)** traditionnelles, conçues pour surveiller les flux réseau, les emails et les transferts de fichiers, se révèlent largement inadaptées aux nouveaux vecteurs de fuite que représentent les LLM. La nature conversationnelle des interactions, le caractère non structuré des données échangées, et la capacité des modèles à transformer et reformuler les informations rendent les approches basées sur le fingerprinting de documents ou le pattern matching simple insuffisantes. En 2026, l'émergence d'une nouvelle génération de solutions **DLP spécifiquement conçues pour l'IA générative** marque un tournant dans la protection des données. Ces solutions combinent l'analyse sémantique, la détection contextuelle et l'intelligence artificielle elle-même pour identifier les fuites de données dans des formats que les outils classiques ne peuvent pas détecter.

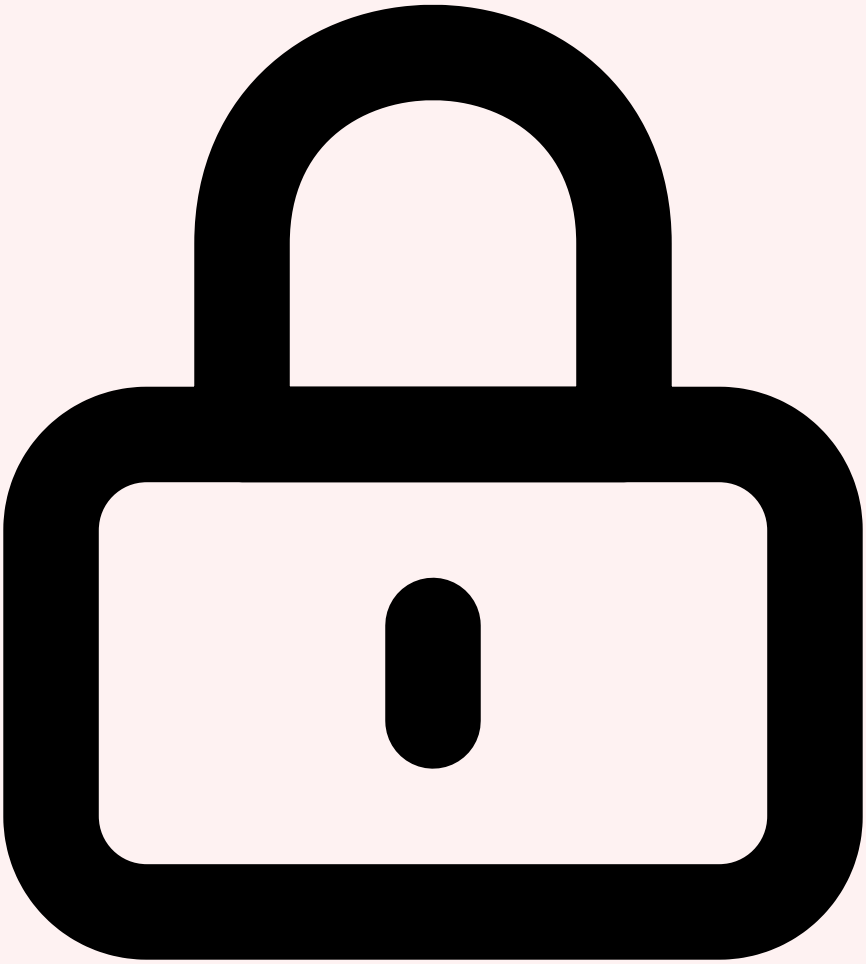


### DLP classique vs DLP pour LLM

Le DLP classique repose sur des **dictionnaires de mots-clés**, des **expressions régulières** et des **fingerprints de documents** pour identifier les données sensibles dans des canaux de communication bien définis (email, web, USB, impression). Cette approche atteint ses limites face aux LLM pour plusieurs raisons fondamentales. Premièrement, les données sensibles ne transitent plus sous forme de fichiers identifiables mais sous forme de **texte libre dans des prompts conversationnels** — un utilisateur peut reformuler les données d'un document classifié sans jamais copier le document lui-même. Deuxièmement, les LLM opèrent via des **API HTTPS chiffrées** qui rendent le DLP réseau aveugle au contenu des requêtes. Troisièmement, la transformation sémantique des données par le LLM (paraphrase, traduction, résumé) élimine les signatures utilisées par le fingerprinting. Enfin, les LLM peuvent **générer des données sensibles** en sortie (via mémorisation ou inférence) — un vecteur de fuite que le DLP traditionnel, focalisé sur les sorties humaines, ne couvre pas. Pour approfondir, consultez Top 10 des Attaques.

## Input Filtering et Output Filtering

L'**input filtering** constitue la première barrière de protection : chaque prompt est intercepté et analysé avant d'être transmis au LLM. Le processus comprend le **PII scanning** (détection et anonymisation des données personnelles via Presidio ou équivalent), le **secrets scanning** (détection de clés API, tokens, mots de passe via des patterns comme ceux de GitGuardian ou TruffleHog), le **content classification** (évaluation du niveau de sensibilité du prompt par rapport à la politique de sécurité), et le **topic restriction** (blocage des requêtes portant sur des sujets interdits comme les données militaires ou les stratégies M&A en cours). L'**output filtering** applique les mêmes contrôles aux réponses du LLM avant leur livraison à l'utilisateur. Ce double filtrage est critique car le LLM peut restituer des données sensibles mémorisées même si le prompt d'entrée était parfaitement sain. L'output filtering ajoute une vérification de **compliance** qui s'assure que la réponse ne contient pas d'informations violant les politiques réglementaires (données de santé sans consentement, données financières non autorisées).

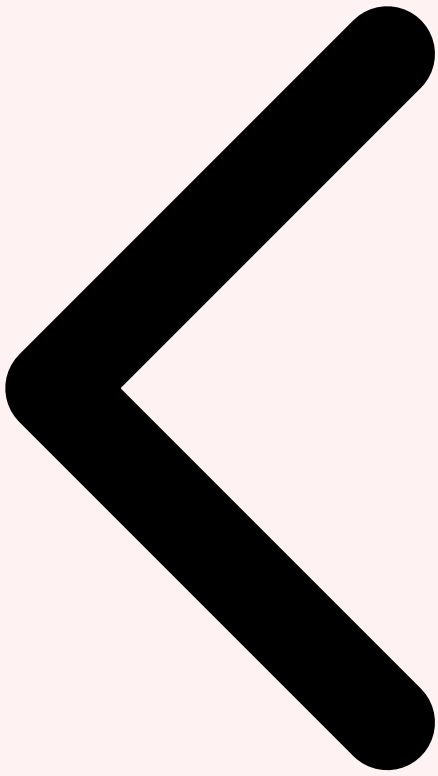


### Guardrails : NeMo Guardrails, LLM Guard, Rebuff

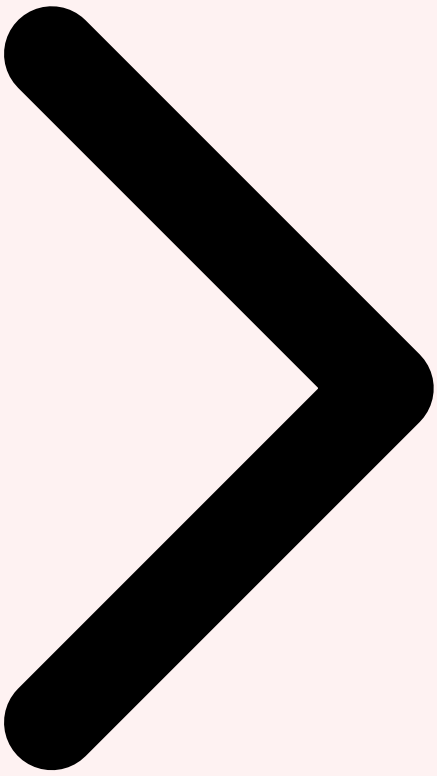
Les frameworks de **guardrails** ajoutent une couche de protection programmatique entre l'utilisateur et le LLM. **NVIDIA NeMo Guardrails** permet de définir des rails de conversation en Colang (un langage déclaratif) qui restreignent les sujets abordables, bloquent les tentatives de prompt injection, et filtrent les réponses contenant des données sensibles. **LLM Guard** (Protect AI) offre un ensemble de scanners prêts à l'emploi pour la détection de PII, de secrets, de contenu toxique et de prompt injections, avec une intégration simple via des API REST. **Rebuff** se spécialise dans la détection et le blocage des tentatives de prompt injection en combinant des heuristiques, des embeddings de similarité et un LLM juge qui évalue si le prompt tente de manipuler le système. En 2026, l'approche recommandée combine ces outils en couches : NeMo Guardrails pour les politiques de haut niveau, LLM Guard pour le scanning technique, et un outil de détection d'injection comme Rebuff ou Lakera Guard pour la protection contre la manipulation.

- ▹**DLP classique insuffisant :** le fingerprinting et le pattern matching simple ne détectent pas les fuites via reformulation, paraphrase ou inférence par le LLM

- ▹**Input + Output filtering :** le scanning bidirectionnel est indispensable — l'input filtering protège les données de l'utilisateur, l'output filtering protège contre la mémorisation du modèle
- ▹**Guardrails combinés :** NeMo Guardrails pour les politiques, LLM Guard pour le scanning technique, Rebuff/Lakera pour la détection d'injection
- ▹**Architecture proxy :** le DLP proxy intercepte toutes les requêtes API vers le LLM, applique les contrôles, et route vers le modèle ou bloque selon la politique

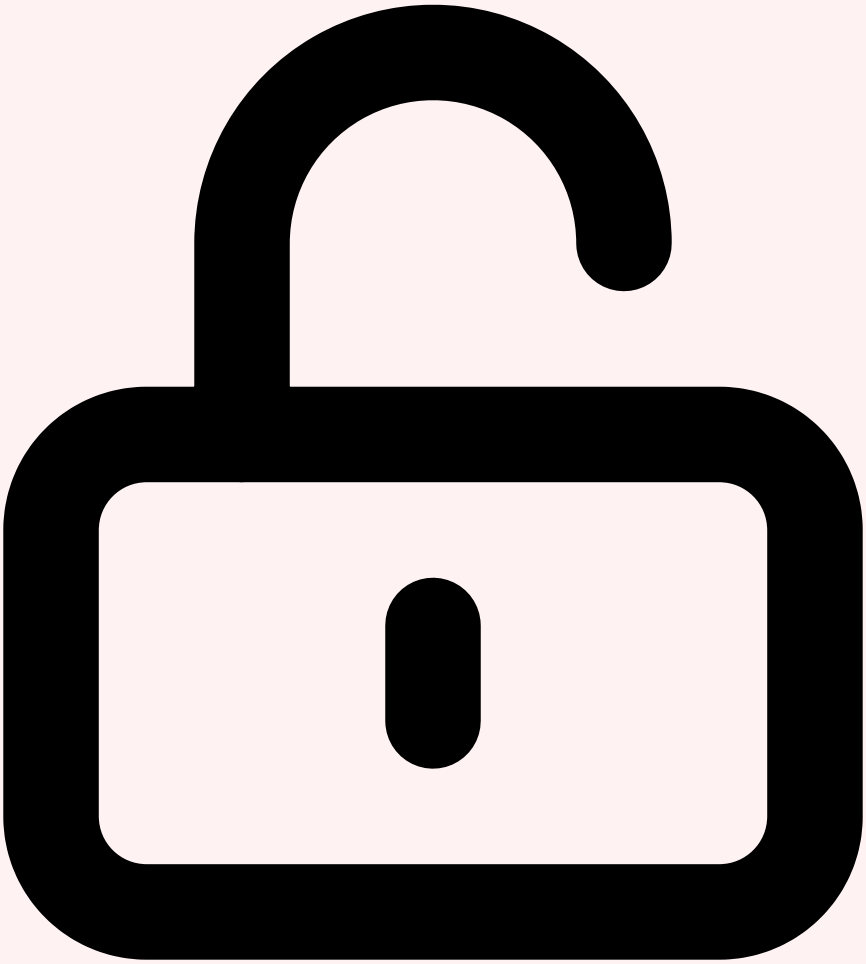


Détection PII Stratégies DLP IA Anonymisation et Privacy



## 5 Techniques d'Anonymisation et de Privacy

Au-delà de la détection et du filtrage des données sensibles, les **techniques d'anonymisation et de privacy-preserving** offrent des approches proactives pour réduire structurellement l'exposition des données dans les systèmes LLM. Ces techniques interviennent à différents niveaux du pipeline : avant l'envoi au LLM (anonymisation des prompts), pendant l'entraînement (differential privacy), et au niveau architectural (federated learning, synthetic data). L'objectif commun est de permettre l'utilisation des LLM pour des tâches à forte valeur ajoutée tout en garantissant que les données personnelles et confidentielles ne sont jamais exposées au modèle sous leur forme originale, ou que le modèle est mathématiquement incapable de les restituer.



### Pseudonymisation réversible vs anonymisation irréversible

La **pseudonymisation** remplace les identifiants directs (nom, email, téléphone) par des tokens artificiels tout en conservant une table de correspondance chiffrée qui permet de restaurer les données originales. Dans le contexte des LLM, cette approche est particulièrement pertinente car elle permet d'envoyer un prompt anonymisé au modèle, puis de réinjecter les données réelles dans la réponse avant de la livrer à l'utilisateur. Par exemple, « Le client Jean Dupont souhaite un contrat » devient « Le client TOKEN_C42 souhaite un contrat » pour le LLM, puis « Jean Dupont » est réinjecté dans la réponse. La table de mapping est stockée en mémoire pendant la durée de la session et détruite ensuite. L'**anonymisation irréversible**, en revanche, détruit définitivement le lien entre le token et la donnée originale. Elle est utilisée lorsque la réversibilité n'est pas nécessaire : logs d'audit, datasets d'évaluation, métriques de performance. Le RGPD distingue clairement ces deux approches : les données pseudonymisées restent des données personnelles soumises au règlement, tandis que les données véritablement anonymisées en sont exclues. Le choix entre les deux dépend donc directement du cas d'usage et des obligations réglementaires applicables.

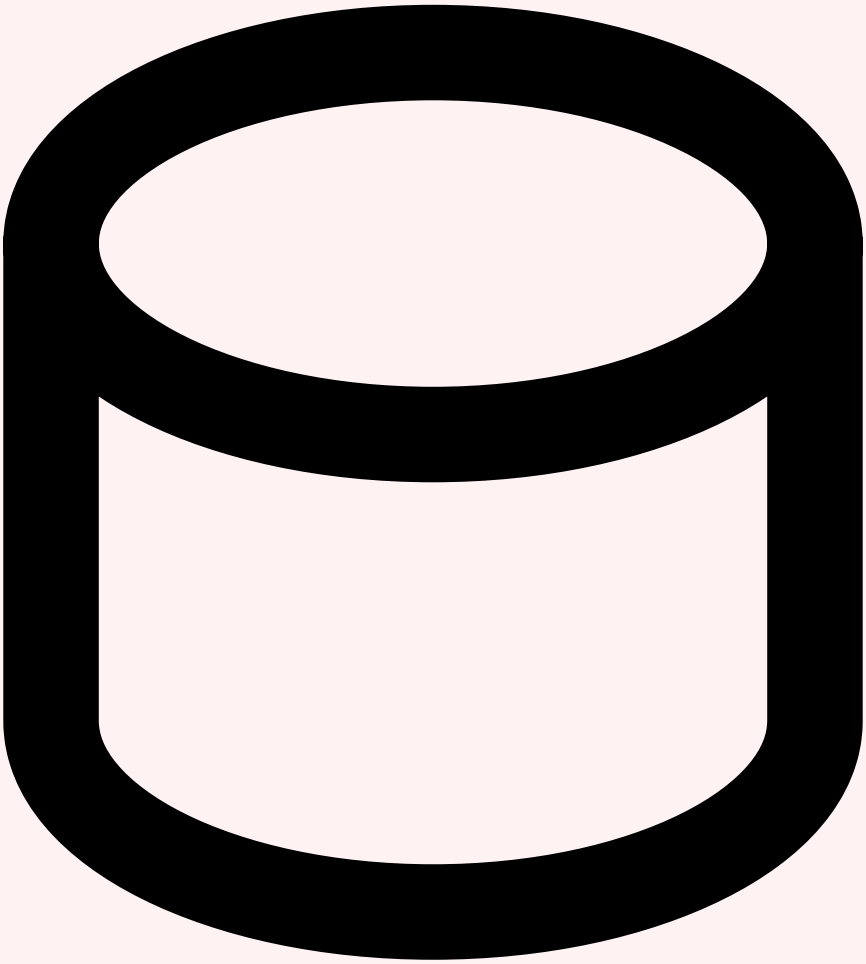
### Tokenisation des PII et mapping réversible

La **tokenisation des PII** avant envoi au LLM est la technique la plus largement déployée en production en 2026. Le processus se décompose en quatre étapes : la **détection** identifie toutes les PII dans le prompt via Presidio ou équivalent ; le **remplacement** substitue chaque PII par un token unique de format cohérent (par exemple, [PERSON_1], [EMAIL_1], [PHONE_1]) ; le **mapping** stocke la correspondance token→valeur_réelle dans un vault éphémère ; et la **dé-tokenisation** réinjecte les valeurs réelles dans la réponse du LLM en remplaçant les tokens. Cette approche présente l'avantage majeur de préserver la cohérence sémantique du prompt — le LLM comprend qu'il s'agit d'une personne, d'un email, etc. — tout en protégeant les données réelles. Les tokens doivent être suffisamment distincts pour ne pas être confondus avec du texte normal, et le format doit être cohérent pour que le LLM les traite correctement dans ses réponses. En pratique, un taux de réversibilité de 95 à 98% est observé sur les réponses structurées, les 2 à 5% restants correspondant aux cas où le LLM reformule la réponse d'une manière qui perd la référence au token original.

## Differential Privacy et Federated Learning

La **differential privacy (DP)** fournit une garantie mathématique que la contribution de chaque échantillon individuel dans le dataset d'entraînement ne peut pas être identifiée dans les sorties du modèle. L'algorithme **DP-SGD (Differentially Private Stochastic Gradient Descent)** modifie le processus d'entraînement en ajoutant un bruit calibré aux gradients à chaque étape, limitant ainsi la quantité d'information que le modèle peut extraire de chaque exemple. Le paramètre epsilon (ε) quantifie le niveau de privacy : un ε faible offre une privacy forte mais dégrade les performances du modèle, tandis qu'un ε élevé préserve les performances mais affaiblit les garanties de privacy. En pratique, les travaux de 2025-2026 ont montré qu'un ε entre 6 et 10 offre un compromis acceptable pour les LLM fine-tunés, avec une dégradation de performance de 2 à 5% mesurée sur les benchmarks standards. Le **federated learning** adopte une approche complémentaire en éliminant la centralisation des données : au lieu d'envoyer les données à un serveur central pour entraîner le modèle, ce sont les **gradients ou les mises à jour du modèle** qui sont partagés, les données restant sur les appareils locaux. Cette technique est particulièrement pertinente pour les organisations multi-sites qui souhaitent fine-tuner un LLM sur des

données sensibles distribuées (hôpitaux, cabinets juridiques, institutions financières) sans jamais centraliser les datasets. La combinaison DP + federated learning offre le niveau de protection le plus élevé disponible aujourd'hui.

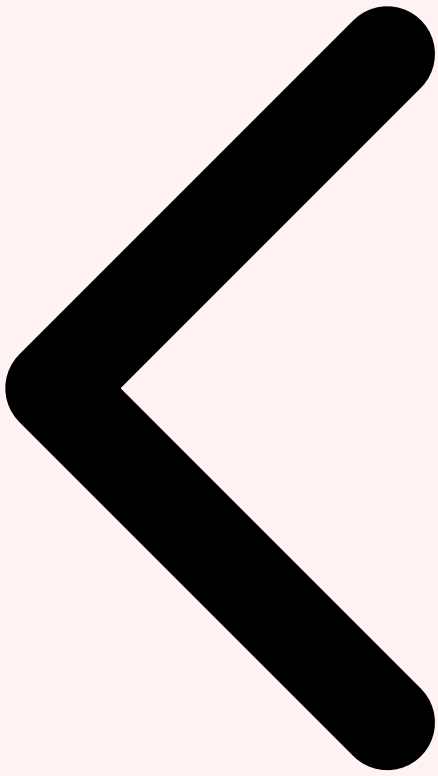


### Synthetic Data Generation

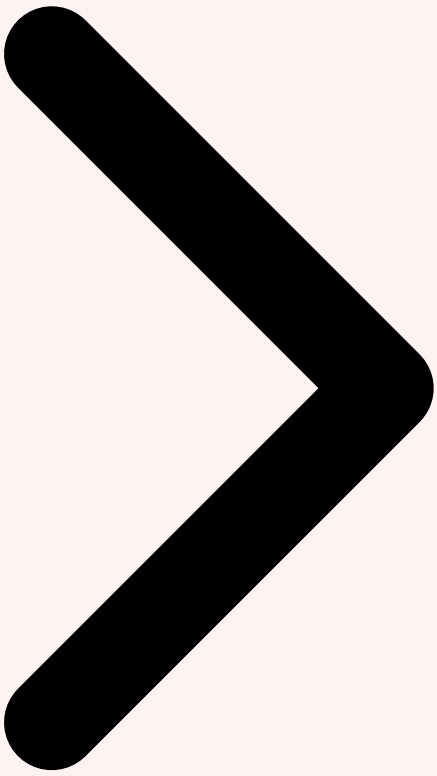
La **génération de données synthétiques** représente l'approche la plus radicale pour éliminer les risques de confidentialité : au lieu de protéger les données réelles, on les remplace entièrement par des données artificielles qui préservent les propriétés statistiques du dataset original sans contenir aucune information réelle. Les modèles génératifs (GANs, VAEs, et désormais les LLM eux-mêmes) produisent des datasets synthétiques réalistes qui peuvent être utilisés pour le fine-tuning, l'évaluation et le développement sans aucun risque de fuite de données personnelles. Des outils comme **Gretel.ai**, **Mostly AI** et **Synthesized** proposent des plateformes de génération de données synthétiques avec des garanties mesurables de privacy (via des métriques comme le singling-out risk et le linkability score). En 2026, les données synthétiques sont de plus en plus utilisées pour les phases de **prototypage et de développement** des applications LLM, réservant les données réelles (avec toutes les protections DLP) aux phases finales de

validation et de production. Cette approche réduit considérablement la surface d'exposition des données sensibles tout au long du cycle de développement. Pour approfondir, consultez Milvus, Qdrant, Weaviate :.

- ▹**Pseudonymisation :** remplacement réversible des PII par des tokens — les données pseudonymisées restent soumises au RGPD, contrairement aux données anonymisées
- ▹**Tokenisation des PII :** technique la plus déployée en production avec un taux de réversibilité de 95-98% — les tokens préservent la cohérence sémantique pour le LLM
- ▹**Differential Privacy :** DP-SGD avec epsilon 6-10 offre le meilleur compromis privacy/performance pour le fine-tuning de LLM — dégradation limitée à 2-5%
- ▹**Données synthétiques :** Gretel.ai, Mostly AI et Synthesized permettent de remplacer les données réelles par des données artificielles statistiquement équivalentes



Stratégies DLP IA Anonymisation et Privacy Conformité RGPD



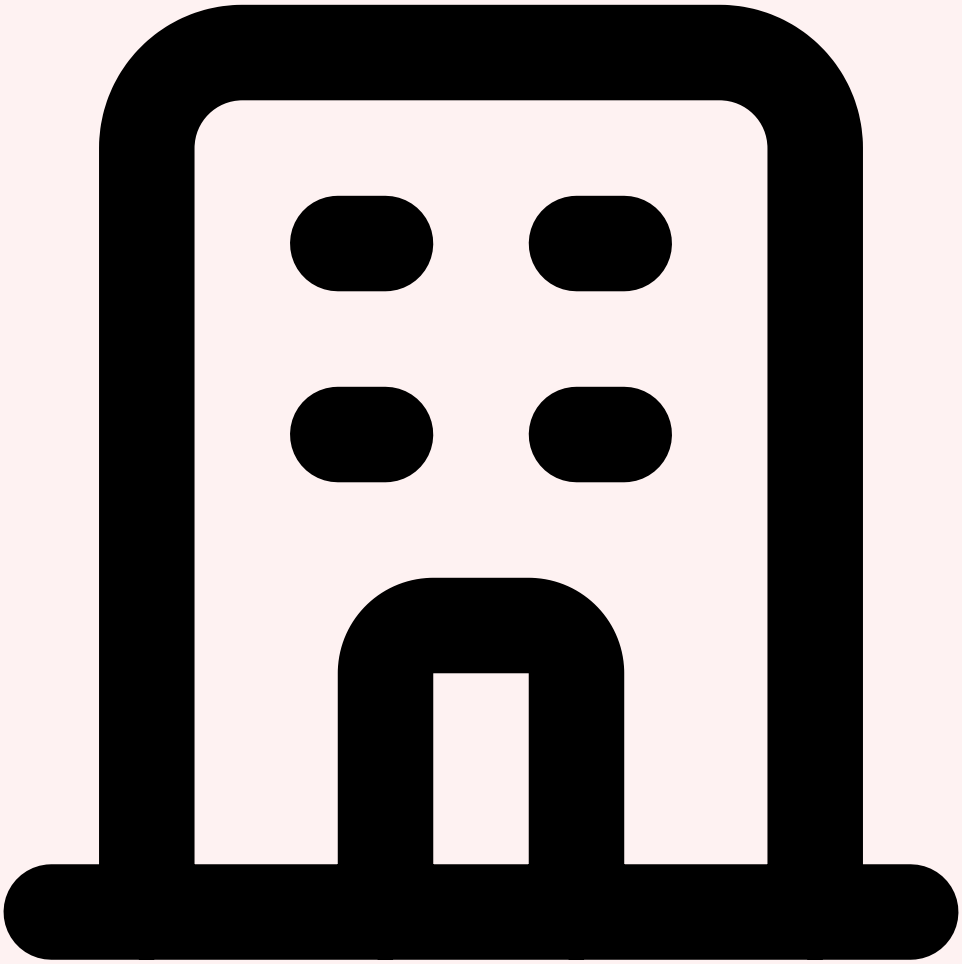
## 6 Conformité RGPD et Réglementaire

L'encadrement réglementaire des **données personnelles dans les systèmes d'IA** s'est considérablement renforcé entre 2024 et 2026, avec l'entrée en application progressive de l'AI Act européen et le durcissement des interprétations du RGPD par les autorités de protection des données. Pour les organisations déployant des LLM, la conformité n'est plus optionnelle mais constitue un prérequis juridique dont le non-respect expose à des sanctions pouvant atteindre 35 millions d'euros ou 7% du chiffre d'affaires mondial pour les violations les plus graves de l'AI Act. La difficulté spécifique des LLM réside dans la multiplicité des traitements de données personnelles qu'ils impliquent — collecte, stockage, traitement automatisé, profilage potentiel, transferts internationaux — souvent de manière opaque et difficilement documentable.

### RGPD appliqué aux LLM

L'application du **RGPD aux Large Language Models** soulève des questions juridiques fondamentales que les régulateurs européens continuent de clarifier. La première question concerne la **base légale du traitement** : l'entraînement d'un LLM sur des données personnelles nécessite une base légale valide (article 6 RGPD). L'intérêt légitime est la base la plus fréquemment invoquée par les fournisseurs de LLM, mais les décisions de la CNIL et du Garante italiano en 2024-2025 ont imposé des conditions strictes : documentation d'une balance des intérêts, mise en œuvre de mécanismes d'opposition facilement accessibles, et limitation de la durée de rétention des données d'entraînement. Le **principe de minimisation** (article 5.1.c) impose de ne collecter et traiter que les données strictement nécessaires — un défi considérable pour les LLM qui sont par nature conçus pour ingérer le maximum de données. Le **droit à l'effacement** (article 17) pose le problème le plus technique : comment supprimer les données d'un individu des poids d'un modèle déjà entraîné ? Les techniques de **machine unlearning** progressent mais restent imparfaites en 2026, et la ré-entraînement complet du modèle après exclusion des données est souvent

économiquement prohibitif. La position pragmatique des régulateurs évolue vers l'acceptation d'une anonymisation effective des sorties comme alternative à l'effacement des poids.



### AI Act : transparence et documentation

L'**AI Act européen**, dont les premières obligations sont entrées en vigueur en février 2025 et les exigences complètes s'appliqueront en août 2026, introduit un cadre de classification des systèmes d'IA par niveau de risque. Les LLM déployés dans des contextes à haut risque (recrutement, crédit, santé, justice) sont soumis à des **obligations de transparence renforcées** : documentation technique complète du système, évaluation des risques incluant les biais et la discrimination, supervision humaine obligatoire, et journalisation des décisions. Pour les **modèles de fondation** (GPAI models), l'AI Act impose des obligations spécifiques aux fournisseurs : documentation des données d'entraînement (incluant les mesures de protection des données personnelles), évaluation et atténuation des risques systémiques, tests de robustesse et de sécurité, et notification des incidents graves. Les fournisseurs de LLM à risque systémique (modèles dépassant 10^25 FLOPS d'entraînement) sont soumis à des **audits obligatoires** et doivent maintenir un système de gestion des risques continu. Pour les organisations utilisatrices de LLM, l'AI Act impose de

s'assurer que leur usage est conforme à la classification de risque, de maintenir la supervision humaine requise, et de documenter les mesures de protection des données personnelles mises en œuvre.

## PCI-DSS, HIPAA et données sectorielles

Les réglementations sectorielles imposent des contraintes supplémentaires spécifiques. **PCI-DSS v4.0** (effective mars 2025) interdit le stockage des données d'authentification sensibles (CVV, PIN, données de bande magnétique) après autorisation — or, un LLM qui reçoit ces données dans un prompt les stocke potentiellement dans ses logs, dans le cache de contexte et, si le modèle est fine-tuné, dans ses poids. La conformité PCI-DSS exige donc un scanning systématique des prompts avec masquage immédiat des données de carte avant tout envoi au LLM, et la purge des logs contenant des PAN (Primary Account Numbers) sous 72 heures. **HIPAA** protège les PHI (Protected Health Information) avec des exigences d'encryption en transit et au repos, de contrôle d'accès granulaire et de journalisation d'audit. L'utilisation d'un LLM cloud pour traiter des données de santé nécessite un **Business Associate Agreement (BAA)** avec le fournisseur — en 2026, seuls

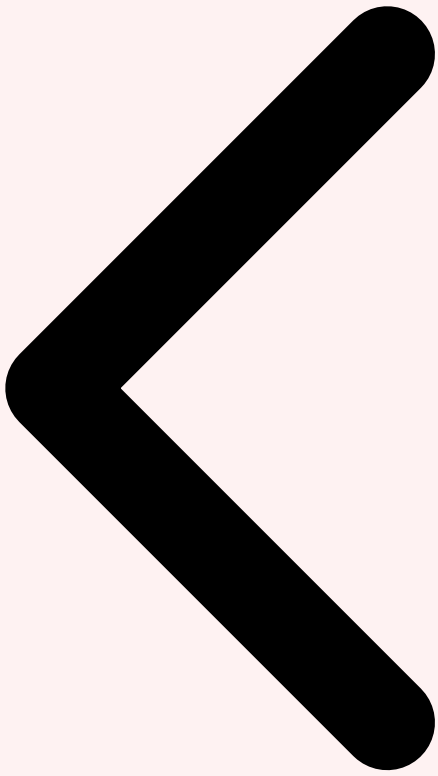
quelques fournisseurs (Azure OpenAI, AWS Bedrock, Google Cloud Vertex AI) proposent des configurations HIPAA-compliant. L'auto-hébergement de modèles open-source (Llama, Mistral) sur infrastructure maîtrisée reste la solution la plus sûre pour les données de santé hautement sensibles.

### DPIA pour les projets LLM

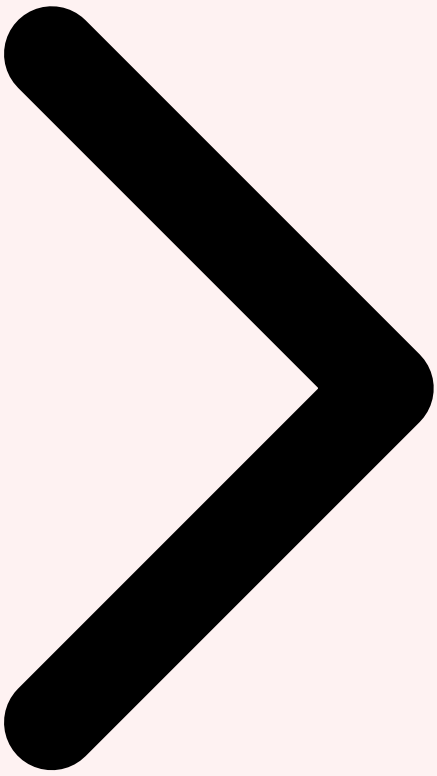
Le **DPIA (Data Protection Impact Assessment)** — ou AIPD (Analyse d'Impact sur la Protection des Données) en français — est obligatoire au titre de l'article 35 du RGPD pour tout traitement « susceptible d'engendrer un risque élevé pour les droits et libertés des personnes physiques ». Les projets LLM cochent systématiquement plusieurs critères déclencheurs : traitement automatisé à grande échelle, évaluation systématique d'aspects personnels (si le LLM prend ou aide des décisions concernant des individus), et utilisation de nouvelles technologies. Le DPIA pour un projet LLM doit documenter : la **description systématique du traitement** (quelles données, quels flux, quels modèles, quels fournisseurs), l'**évaluation de la nécessité et de la proportionnalité** (pourquoi un LLM est nécessaire, pourquoi ces données sont nécessaires), l'**évaluation des risques** pour les personnes concernées (fuite de données, discrimination, décisions automatisées erronées), et les **mesures d'atténuation** prévues (DLP, anonymisation, contrôle d'accès, audit). La

CNIL recommande de réaliser le DPIA avant le déploiement et de le mettre à jour à chaque évolution significative du système, notamment les changements de modèle, les modifications des flux de données, ou les nouveaux cas d'usage.

- ▹**RGPD :** base légale obligatoire, minimisation des données, droit à l'effacement (machine unlearning) — les données pseudonymisées restent soumises au règlement
- ▹**AI Act :** classification par niveau de risque, obligations de transparence et documentation, audits obligatoires pour les modèles de fondation à risque systémique
- ▹**PCI-DSS / HIPAA :** scanning obligatoire des données de carte et de santé avant envoi au LLM — seuls quelques fournisseurs cloud proposent des configurations conformes
- ▹**DPIA :** obligatoire pour tout projet LLM traitant des données personnelles — doit être réalisé avant le déploiement et mis à jour à chaque évolution

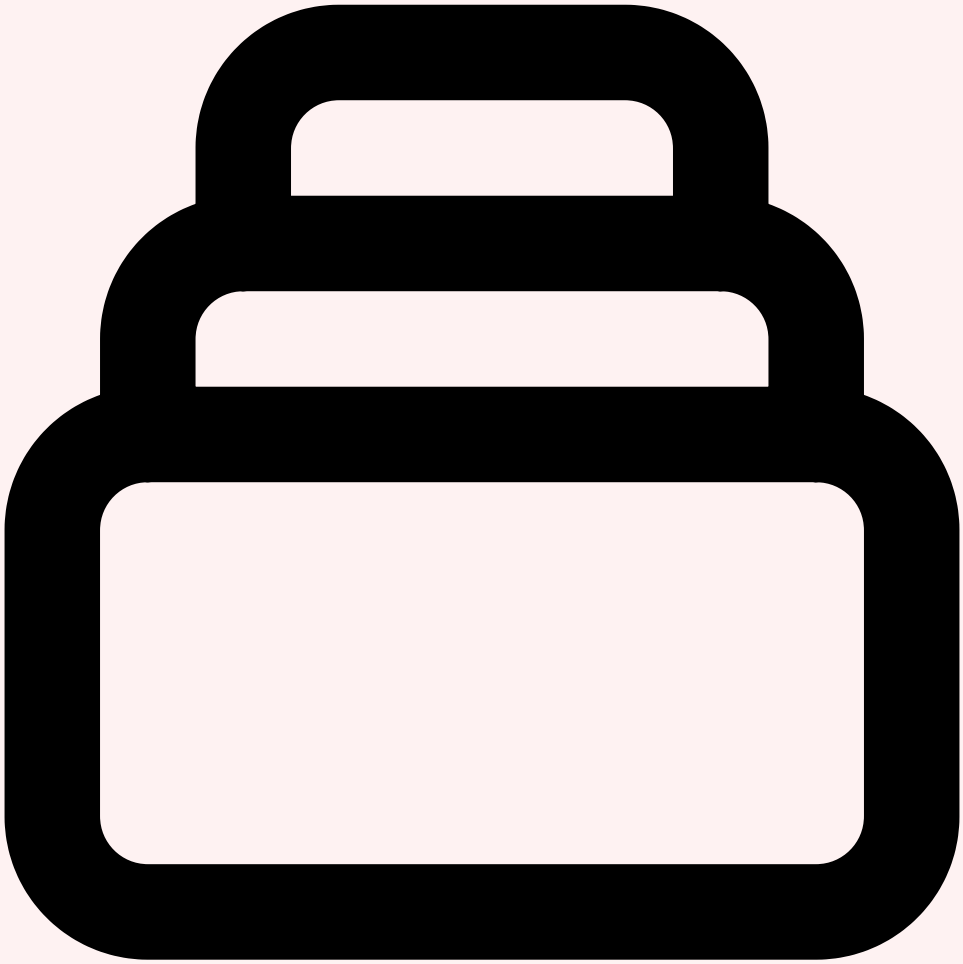


Anonymisation et Privacy Conformité RGPD Implémentation Pratique



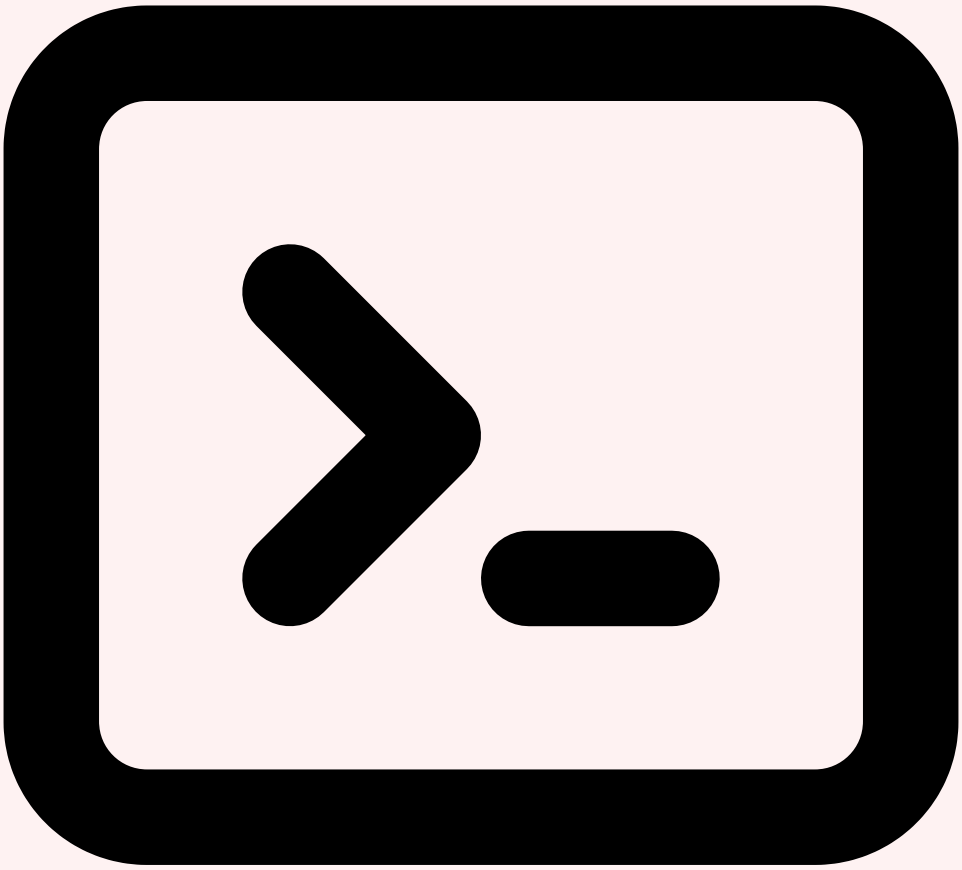
## 7 Implémentation Pratique : Pipeline DLP LLM

La mise en œuvre d'un **pipeline DLP complet pour les applications LLM** nécessite une approche architecturale structurée qui intègre la détection, la protection, le monitoring et la gouvernance dans un système cohérent. Cette section présente une architecture de référence déployable en production, les patterns d'intégration avec les API gateways existantes, les métriques de performance clés, et une checklist opérationnelle pour les RSSI. L'objectif est de fournir un guide actionnable permettant aux équipes sécurité de déployer une protection DLP fonctionnelle en quelques semaines, puis de l'affiner progressivement en fonction des retours de production.



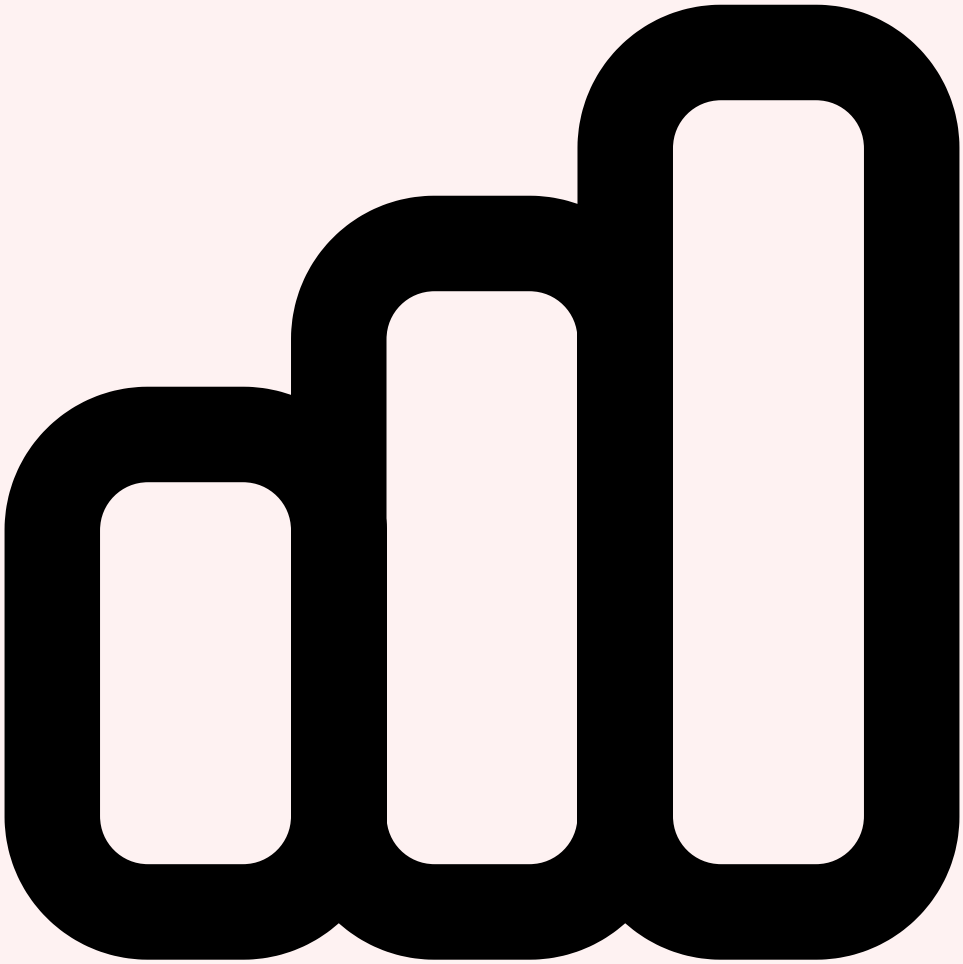
### Architecture de référence DLP pour LLM

L'architecture de référence s'articule autour d'un **proxy DLP** positionné entre les clients (applications, utilisateurs) et les endpoints LLM (API OpenAI, Anthropic, modèles auto-hébergés). Ce proxy intercepte toutes les requêtes et réponses, applique les politiques de sécurité, et route le trafic vers les modèles autorisés. Le proxy se compose de plusieurs modules orchestrés : le **PII Scanner** (Presidio + recognizers custom) détecte et anonymise les données personnelles ; le **Secrets Scanner** (patterns GitGuardian/TruffleHog) identifie les clés API, tokens et credentials ; le **Content Classifier** (modèle de classification fine-tuné) évalue le niveau de sensibilité du contenu ; le **Policy Engine** (OPA ou moteur de règles custom) applique les politiques d'autorisation/blocage ; et le **Audit Logger** enregistre chaque décision dans un store immuable pour la traçabilité. En production, ce proxy est déployé comme un **sidecar container** dans un cluster Kubernetes, avec auto-scaling horizontal pour absorber les pics de charge. La latence ajoutée typique est de 30 à 80 millisecondes par requête, ce qui est acceptable pour la plupart des cas d'usage conversationnels où l'inférence LLM prend 500 ms à 3 secondes.



### Intégration avec les API Gateways

L'intégration du pipeline DLP avec les **API gateways** existants (Kong, Apigee, AWS API Gateway, Azure API Management) permet de capitaliser sur l'infrastructure de gestion d'API déjà en place. L'approche recommandée utilise le pattern **plugin/middleware** : un plugin custom installé sur le gateway intercepte les requêtes vers les endpoints LLM, appelle le service DLP via gRPC ou REST pour le scanning, et bloque ou modifie la requête selon le résultat. Pour **Kong**, le plugin Lua ou Go s'insère dans la phase access/body_filter du cycle de requête. Pour **Apigee**, une policy chain avec callout vers le service DLP est configurée dans le proxy API. L'avantage de cette approche est qu'elle centralise le contrôle DLP au point d'entrée unique de l'API, capturant toutes les requêtes y compris celles provenant d'applications internes, de scripts automatisés et d'intégrations tierces. Les **LLM gateways spécialisés** comme Portkey, LiteLLM et Helicone offrent des intégrations DLP natives qui simplifient le déploiement pour les équipes qui ne disposent pas d'un API gateway existant. Ces gateways spécialisés ajoutent la gestion des clés API LLM, le load balancing entre modèles, le fallback automatique et le caching sémantique en plus des fonctions DLP. Pour approfondir, consultez Reinforcement Learning Appliqué à la Cybersécurité.



### Monitoring, métriques et alerting

Le monitoring du pipeline DLP repose sur quatre catégories de **métriques opérationnelles** essentielles. Le **taux de détection** (true positive rate) mesure le pourcentage de PII et de données sensibles correctement identifiées — la cible est supérieure à 95% pour les PII directs (emails, téléphones, SSN) et supérieure à 85% pour les PII indirects (noms, adresses). Le **taux de faux positifs** (false positive rate) mesure les détections erronées — la cible est inférieure à 5% pour éviter le « DLP fatigue » où les utilisateurs contournent le système en raison de blocages abusifs. La **latence ajoutée** par le pipeline DLP doit rester sous 100 ms au P95 pour ne pas impacter l'expérience utilisateur — un monitoring par percentile (P50, P90, P95, P99) est indispensable pour détecter les dégradations de performance. Le **volume de violations** par catégorie (PII, secrets, compliance) et par source (utilisateur, application, département) fournit la visibilité nécessaire pour identifier les zones à risque et ajuster les politiques. Les alertes doivent être configurées avec des seuils progressifs : notification informative pour les détections isolées, alerte SOC pour les patterns suspects (même utilisateur, nombreuses détections en série), et escalade RSSI pour les violations de données réglementées (données de santé, données de carte).

## Checklist RSSI pour la confidentialité des LLM

| Domaine | Action | Priorité | Outil |
|---|---|---|---|
| Inventaire | Cartographier tous les usages LLM (autorisés et shadow AI) | CRITIQUE | CASB, proxy logs |
| Politique | Définir la politique d'usage acceptable des LLM | CRITIQUE | PSSI, charte IA |
| DLP Input | Déployer le scanning PII/ secrets sur tous les prompts | CRITIQUE | Presidio, LLM Guard |
| DLP Output | Scanner les réponses LLM avant livraison utilisateur | ÉLEVÉ | NeMo Guardrails |
| Audit | Implémenter la journalisation complète des interactions LLM | ÉLEVÉ | ELK, Splunk, Datadog |
| Conformité | Réaliser un DPIA pour chaque projet LLM avec données personnelles | ÉLEVÉ | Template CNIL |
| Anonymisation | Appliquer la tokenisation réversible des PII | MOYEN | Presidio Anonymizer |
| Formation | Sensibiliser les collaborateurs aux risques de fuite via LLM | MOYEN | e-learning, workshops |
| Architecture | Évaluer le self-hosting pour les données les plus sensibles | MOYEN | Llama, Mistral, vLLM |
| Incident | Préparer un plan de réponse aux incidents de fuite de données via LLM | MOYEN | PRI, playbooks SOC |

Cette checklist constitue un cadre opérationnel pour les RSSI et les équipes sécurité chargées de la mise en conformité des projets LLM. Les actions critiques (inventaire, politique, DLP input) doivent être implémentées en priorité absolue, car elles couvrent les risques les plus immédiats et les plus fréquents. Les actions de niveau élevé (DLP output, audit, DPIA) constituent la deuxième vague de déploiement, typiquement dans les 2 à 3 mois suivant le lancement. Les actions de niveau moyen (anonymisation, formation, architecture, incident) complètent le dispositif de protection et doivent être planifiées dans les 6 mois. L'ensemble du dispositif doit être revu trimestriellement pour intégrer les nouvelles menaces, les évolutions réglementaires et les retours d'expérience opérationnels. L'indicateur de maturité le plus pertinent est le **pourcentage de requêtes LLM passant par le pipeline DLP** — l'objectif est d'atteindre 100% des usages autorisés dans les 3 mois et de réduire le Shadow AI non couvert à moins de 5% dans les 6 mois.

- ▹**Proxy DLP :** architecture centralisée interceptant toutes les requêtes LLM — latence de 30-80ms acceptable pour les cas d'usage conversationnels
- ▹**API Gateway :** intégration via plugins Kong/Apigee ou LLM gateways spécialisés (Portkey, LiteLLM) pour capitaliser sur l'infrastructure existante
- ▹**Métriques clés :** taux de détection > 95% (PII directs), faux positifs < 5%, latence < 100ms P95, couverture 100% des flux autorisés
- ▹**Maturité :** déploiement en 3 vagues — critique (0-1 mois), élevé (1-3 mois), moyen (3-6 mois) — avec revue trimestrielle du dispositif

### Besoin d'un accompagnement expert ?

Nos consultants en cybersécurité et IA vous accompagnent dans vos projets. Devis personnalisé sous 24h.

### Références et ressources externes

- OWASP LLM Top 10 — Les 10 risques majeurs pour les applications LLM
- MITRE ATLAS — Framework de menaces pour les systèmes d'intelligence artificielle
- NIST AI RMF — AI Risk Management Framework du NIST
- arXiv — Archive ouverte de publications scientifiques en IA
- HuggingFace Docs — Documentation de référence pour les modèles de ML

Pour approfondir ce sujet, consultez notre outil open-source llm-security-scanner qui facilite l'audit de sécurité des modèles de langage.

**Sources et références :** ArXiv IA · Hugging Face Papers

## FAQ

### Qu'est-ce que Confidentialité des Données dans les LLM ?

Le concept de Confidentialité des Données dans les LLM est détaillé dans les premières sections de cet article, qui couvrent les fondamentaux, les enjeux et le contexte opérationnel. Pour un accompagnement sur ce sujet, contactez nos experts.

### Pourquoi Confidentialité des Données dans les LLM est-il important en cybersécurité ?

La compréhension de Confidentialité des Données dans les LLM permet aux équipes de sécurité d'améliorer leur posture défensive. Les sections « Table des Matières » et « 2 Typologie des Données Sensibles dans les LLM » détaillent les raisons de cette importance. Pour un accompagnement sur ce sujet, contactez nos experts.

### Comment mettre en œuvre les recommandations de cet article ?

Les recommandations pratiques sont détaillées tout au long de l'article, avec des commandes, des outils et des méthodologies éprouvées. La section « Conclusion » fournit une synthèse actionnable. Pour un accompagnement sur ce sujet, contactez nos experts.

## Conclusion

Cet article a couvert les aspects essentiels de Table des Matières, 1 Les Risques de Confidentialité des LLM, 2 Typologie des Données Sensibles dans les LLM. La mise en pratique de ces recommandations permet de renforcer significativement la posture de securite de votre organisation.

**Ayi NEDJIMI Consultants** — Expert cybersécurité offensive & intelligence artificielle

ayinedjimi-consultants.fr · ayi@ayinedjimi-consultants.fr