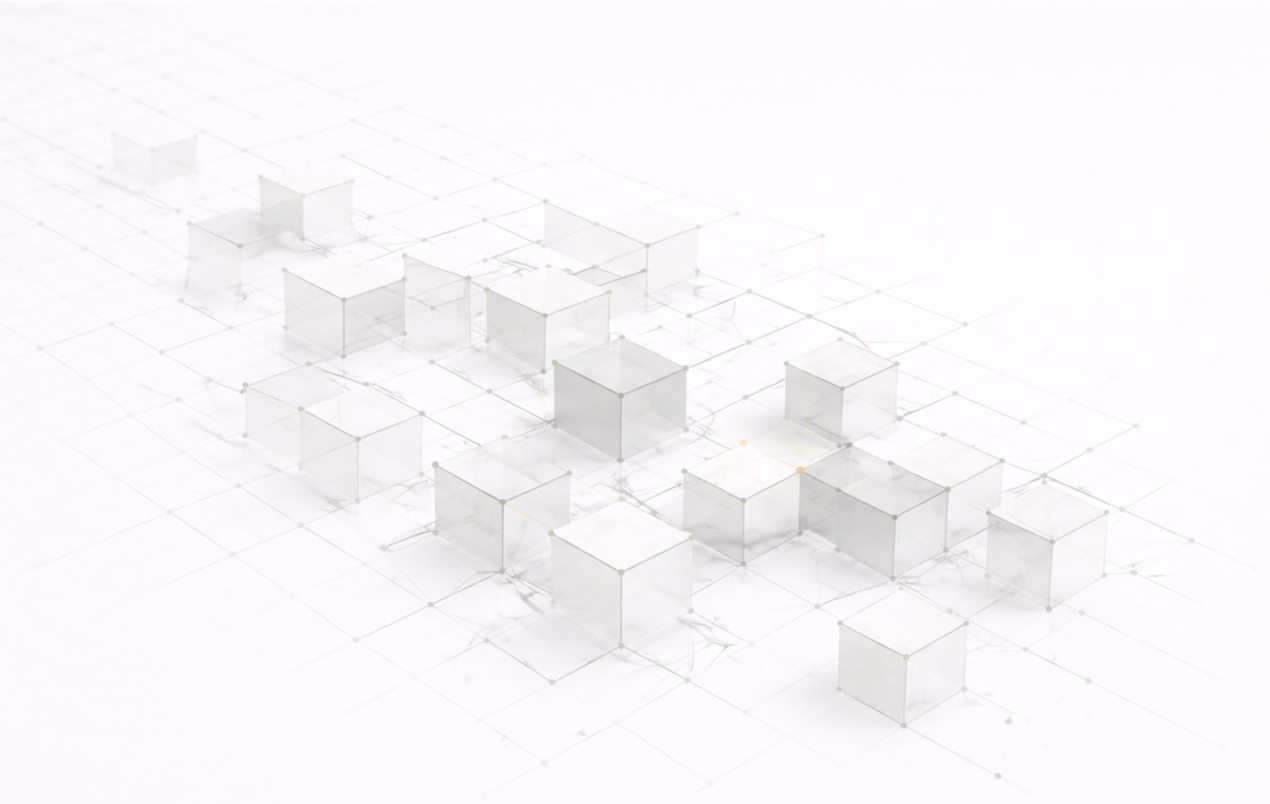

APOLLO Data Auditor

LIVRE BLANC · QUALITÉ & IA

Vos données sont-elles prêtes pour l'IA ?

Pourquoi 85% des projets IA échouent avant même que le modèle ne soit un problème

Module Qualité & IA — Avril 2026

RÉSUMÉ EXÉCUTIF

95% des pilotes GenAI ne produisent aucun retour sur investissement mesurable — malgré 30–40 milliards de dollars en dépenses mondiales en IA (MIT, 2025).

L'approche dominante est centrée sur l'IA : choisir un modèle, acheter une licence, lancer un pilote — et espérer que les données suivront. Rarement. Gartner attribue 85% des défaillances de l'IA à la qualité des données, pas à la qualité du modèle. La Rand Corporation a trouvé que les projets IA échouent deux fois plus que les projets informatiques traditionnels.

Ce document examine pourquoi les défaillances de l'IA sont des défaillances de données, ce qu'une approche centrée sur les données change, et pourquoi — avec la loi européenne sur l'IA entrant en vigueur le 2 août 2026 — la question n'est plus optionnelle.

1. LE PROJET IA QUI N'A JAMAIS ÉTÉ EXPÉDIÉ

Selon le rapport TechShift 2026 Enterprise AI Readiness Report, 65% des initiatives d'IA qui se sont figées en 2024–2025 ont cité la qualité des données ou l'accessibilité comme bloqueur principal — pas la capacité du modèle, pas le budget. Un schéma récurrente traverse les industries : une entreprise embauche des data scientists, sélectionne un modèle, lance un pilote — et observe 80% du temps de l'équipe disparaître dans le nettoyage des données, la déduplification et la réconciliation de schémas. Dix-huit mois plus tard, le projet est abandonné, trois fois le budget dépassé, sans résultat utilisable.

Le modèle n'était pas le problème. Les données étaient. Et dans la plupart des cas, personne n'avait vérifié avant que le budget ne soit engagé.

Ce schéma se répète à grande échelle. La Rand Corporation a constaté que plus de 80% des projets d'IA échouent — deux fois le taux des projets informatiques traditionnels. Les chercheurs du MIT ont confirmé en 2025 que 95% des pilotes GenAI n'ont produit aucun ROI mesurable — malgré des dizaines de milliards en dépenses mondiales. La cause en est structurelle : les organisations investissent dans les modèles avant d'auditer leurs données. Elles demandent « quelle IA devrions-nous déployer ? » avant de demander « nos données sont-elles prêtes pour une IA quelconque ? »

APOLLO™ DATA AUDITOR EN ACTION — TROIS PROJETS IA SCORÉS AVANT LANCEMENT

Chaque cas ci-dessous est un projet d'IA qui a échoué — non pas à cause du modèle, mais à cause des données. Voici ce qu'un scan APOLLO™ Data Auditor aurait révélé avant le premier sprint.

Cas 1 — Société de services financiers européenne, projet de scoring crédit IA abandonné après 14 mois

Profil de scan : 3 bases de données, 680 000 dossiers clients. Types de données : relevés de revenus, historiques de transactions, demandes de crédit. Problème : 34% de doublons sur 3 systèmes sources, 22% de valeurs manquantes dans les champs clés (revenus, statut professionnel). Incohérence de schéma : formats de date dans 4 standards différents.

SCORE	VALEUR	STATUT
Maturité IA (S012)	**27 / 100**	NON PRÊT
Qualité des Données	31 / 100	CRITIQUE
Gouvernance des Données	24 / 100	CRITIQUE
Dépassement de budget estimé	**480 000 €**	—

DIMENSION	SCORE	CONSTAT
Complétude	18 / 100	22% de valeurs manquantes dans les champs revenus/emploi — variables clés du modèle crédit
Unicité	24 / 100	34% de doublons entre sources — le modèle s'entraînerait sur des données contaminées
Cohérence	31 / 100	4 formats de date, 3 représentations monétaires, 2 conventions de nommage
AI Act Art. 10	15 / 100	Gouvernance des données d'entraînement non documentée — non conforme pour IA à risque élevé

PRIORITÉ	ACTION	IMPACT ESTIMÉ
P1	Dédupliquer 231 000 enregistrements avant le gel des données d'entraînement	Supprime la contamination — Maturité IA de 27 à 54+
P1	Standardiser le schéma sur les 3 systèmes sources — date, devise, nommage	Le problème de complétude ne peut être résolu sans alignement du schéma
P1	Documenter la gouvernance des données d'entraînement (conformité Art. 10)	Requis avant l'entrée en vigueur de l'AI Act — échéance août 2026

> *Le projet a été abandonné après 14 mois. Le score de Maturité IA APOLLO aurait été de 27/100 dès le premier jour. Le dépassement de budget estimé : 480 000 €. Le temps pour corriger les problèmes de données avant de commencer : 8 semaines. L'équipe projet n'avait jamais réalisé d'audit de données avant le premier sprint.*

Cas 2 — Réseau de santé américain, assistant IA de diagnostic bloqué par double conformité HIPAA + AI Act

Profil de scan : 2 bases de données, 4,2 millions de dossiers patients. Données d'entraînement pour IA diagnostique. Problème : 18% des dossiers contiennent des données de personnes décédées depuis plus de 10 ans (pas de processus de suppression), 12% de doublons issus de migrations. Catégorie à risque élevé AI Act : diagnostic médical. Conformité HIPAA : non vérifiée sur les données d'entraînement.

SCORE	VALEUR	STATUT
Maturité IA (S012)	**19 / 100**	NON PRÊT
Conformité HIPAA/AI Act	Note F	CRITIQUE
Gouvernance des Données	14 / 100	CRITIQUE
Exposition réglementaire estimée	**2 100 000 $**	—

DIMENSION	SCORE	CONSTAT
Minimisation des données	8 / 100	756 000 dossiers sans base de conservation — patients décédés, pas de conservation légale
Qualité données d'entraînement	22 / 100	504 000 doublons gonflant le taux de faux positifs à l'entraînement
AI Act Art. 10	0 / 100	Pas de documentation sur la gouvernance des données d'entraînement. IA à risque élevé. Obligation.
AI Act Art. 15	12 / 100	Pas de posture cybersécurité documentée pour le système d'IA

PRIORITÉ	ACTION	IMPACT ESTIMÉ
P1	Supprimer les 756 000 dossiers sans base de conservation avant le gel des données	Supprime la violation HIPAA + contamination Art. 10 AI Act
P1	Dédupliquer 504 000 enregistrements — documenter la méthodologie pour Art. 10	Qualité données d'entraînement de 22 à 68+
P1	Créer la documentation de gouvernance des données d'entraînement (Art. 10 — obligatoire IA risque élevé)	Requis — AI Act en vigueur le 2 août 2026

> *L'IA diagnostique était classée à risque élevé selon l'article 6 de l'AI Act. Sans conformité Art. 10, le déploiement dans l'UE est illégal après août 2026. Score APOLLO : 0/100 sur AI Act Art. 10. Les 756 000 dossiers sans base de conservation auraient créé des violations simultanées HIPAA et AI Act dans tout modèle déployé.*

Cas 3 — Groupe de distribution français, moteur de recommandation client produisant des résultats discriminatoires

Profil de scan : base de données CRM, 1,8 million de dossiers clients. Moteur de recommandation IA entraîné sur 3 ans d'historique d'achats. Problème : historique biaisé par la période COVID (2020-2022) — 40% des données d'entraînement non représentatives. Attributs sensibles (proxy d'âge, code postal) présents sans revue de biais. RGPD Art. 22 (décision automatisée) : pas de documentation.

SCORE	VALEUR	STATUT
Maturité IA (S012)	**34 / 100**	NON PRÊT
Qualité des Données	41 / 100	À AMÉLIORER
Conformité RGPD	28 / 100	Note F
Exposition réglementaire estimée	**340 000 €**	—

DIMENSION	SCORE	CONSTAT
Représentativité	12 / 100	40% des données d'entraînement issues de la période anomalie COVID — statistiquement non représentatives
Indicateurs de biais	8 / 100	Proxy d'âge + code postal (indicateur socio-économique) dans les variables — non documenté
RGPD Art. 22	0 / 100	Prise de décision automatisée déployée sans documentation de transparence
Fraîcheur	45 / 100	Données d'entraînement vieilles de 3 ans pour un modèle générant des recommandations en temps réel

PRIORITÉ	ACTION	IMPACT ESTIMÉ
P1	Exclure les données 2020-2022 du jeu d'entraînement — marquer comme non représentatives	Représentativité : de 12 à 68+
P1	Documenter la logique de décision automatisée (RGPD Art. 22)	Requis — actuellement 0% conforme
P2	Supprimer le proxy d'âge et le code postal des variables ou documenter la revue de biais	Élimine le risque de résultats discriminatoires

> *Le modèle était en production depuis 18 mois avant qu'un audit de biais ne détecte le problème des données COVID. Conformité RGPD Art. 22 : 0/100. La correction — exclure 40% des données d'entraînement et ré-entraîner — a pris 6 semaines. APOLLO aurait signalé l'anomalie COVID et l'écart Art. 22 avant le déploiement du premier modèle.*

2. LES TROIS PROBLÈMES DE DONNÉES QUE L'IA NE PEUT PAS RÉSOUDRE POUR VOUS

L'IA ne corrige pas les mauvaises données. Elle les amplifie. Trois catégories de problèmes de données bloquent systématiquement ou corrompent les déploiements d'IA — et aucune d'elles ne sont visibles sans audit.

Problème 1 : Vos données sont contaminées.

Une entreprise entraîne un chatbot de support interne sur un corpus de tickets clients. Les tickets n'ont jamais été anonymisés. Le modèle mémorise et régurgite les données personnelles — adresses e-mail, numéros de téléphone, adresses physiques — dans ses réponses. Le rapport International AI Safety Report 2025 identifie la suppression des données d'identité des données d'entraînement comme l'un des principaux défis non résolus du déploiement de l'IA.

Un outil de recrutement entraîné sur dix ans de données d'embauche reproduit les biais codés dans cet historique — favorisant certains profils, en pénalisant d'autres. Les données n'ont pas été auditées pour leur représentativité avant l'entraînement. Le résultat : la discrimination systémique et l'exposition réglementaire en vertu de l'annexe III de la loi sur l'IA (systèmes d'emploi = catégorie à haut risque).

Problème 2 : Vos employés utilisent déjà l'IA — sans vous.

Le rapport IBM Cost of a Data Breach 2025 a constaté que 20% des brèches impliquent maintenant de l'IA fantôme — les employés utilisant des outils d'IA non autorisés avec les données de l'entreprise. Le coût moyen d'une brèche d'IA fantôme est de 4,63 millions de dollars, par rapport à 3,96 millions de dollars pour une brèche standard — une prime de 670 000 $.

63% des organisations n'ont pas de politique de gouvernance de l'IA. 97% des organisations qui ont connu une brèche liée à l'IA n'avaient aucun contrôle d'accès à l'IA dédié. Seulement 37% ont un mécanisme pour détecter ou approuver l'utilisation d'outils d'IA.

Les employés collent les données des clients dans ChatGPT pour rédiger des propositions. Ils téléchargent des feuilles de calcul sur des assistants IA. Ils alimentent des documents confidentiels dans des outils qui stockent les entrées sur des serveurs qu'ils ne contrôlent pas. Ce n'est pas malveillant — c'est l'absence de gouvernance rencontrant la disponibilité d'outils puissants.

Problème 3 : La loi sur l'IA arrive — et votre gouvernance des données n'est pas prête.

La loi européenne sur l'IA entre en vigueur le 2 août 2026 pour les systèmes d'IA à haut risque (annexe III : emploi, notation de crédit, santé, éducation, biométrie, infrastructure critique). Pénalités : jusqu'à 35 millions d'€ ou 7% du chiffre d'affaires mondial.

L'article 10 exige une gouvernance formelle des données pour les ensembles de données d'entraînement — qualité, complétude, représentativité, traçabilité. L'article 15 exige des mesures de cybersécurité proportionnées au niveau de risque du système d'IA.

La plupart des outils de conformité pour la loi sur l'IA sont déclaratifs — listes de contrôle et questionnaires. Mais l'article 10 ne peut pas être satisfait par déclaration. Il nécessite des métriques mesurables sur les données réelles : scores de complétude, taux de cohérence, stabilité de schéma, niveaux de contamination aux données d'identité. Aucun questionnaire ne produit ces chiffres.

3. À QUOI RESSEMBLE UNE APPROCHE CENTRÉE SUR LES DONNÉES

L'alternative à l'approche centrée sur l'IA n'est pas « pas d'IA ». C'est l'approche centrée sur les données : auditer les données avant d'investir dans le modèle. Mesurer la préparation avant de s'engager dans le budget. Identifier les blocages avant que les data scientists passent 18 mois à les découvrir manuellement.

Un audit centré sur les données répond à quatre questions :

La qualité de vos données est-elle suffisante pour l'IA ? Complétude, cohérence, fraîcheur — mesurées par source, par table, par champ. Pas « nos données sont généralement bonnes » mais « cette base de données a 62% de complétude, la stabilité du schéma est 1/5, et 6% des fichiers sont dans des formats exploitables par les LLM. »

Qu'est-ce qui bloque le déploiement de l'IA en ce moment ? Blocages concrets et actionnables — pas des catégories de risque abstraites. « 2 fichiers contenant des données d'identité non chiffrées sont accessibles par Copilot. » « 2 tables de base de données n'ont pas de clé primaire — l'anonymisation est structurellement impossible. » « L'extractabilité est de 6% — vos données ne sont pas utilisables par aucun LLM sans transformation. »

Quel est votre niveau de maturité des données ? Dix dimensions notées de 1 à 5 : conventions de nommage, documentation, chiffrement, gestion du cycle de vie, cohérence, qualité de schéma, piste d'audit, classification, contrôle d'accès, sensibilisation aux coûts. Niveau actuel par rapport à l'objectif. « Vous êtes au niveau 2. Le niveau 3 (Proactif) nécessite d'améliorer la documentation de 1/5 à 3/5 et le chiffrement de 1/5 à 3/5. »

Êtes-vous exposé en vertu de la loi sur l'IA ? Si des signaux d'IA sont détectés dans votre environnement, le module les signale et fournit des métriques de gouvernance des données de l'article 10 et des notes de posture cybersécurité de l'article 15. Si aucun signal d'IA n'est détecté, la réponse est honnête : « Loi sur l'IA non applicable à ce stade. » Pas de survalorisation. Pas d'urgence fabriquée.

---

4. COMMENT APOLLO DATA AUDITOR NOTE LA PRÉPARATION À L'IA

Les quatre questions ci-dessus existent aujourd'hui dans le module Intelligence. Voici ce qu'il produit à partir d'un scan réel.

Vos données restent les vôtres. Un agent natif s'exécute localement sur Windows, Linux et macOS (arm64). Il analyse les fichiers, les bases de données, le stockage cloud, Active Directory et l'infrastructure. Seules les métadonnées et les compteurs transitent vers le tableau de bord cloud. L'agent est open source (BSL 1.1) — vérifiable sur GitHub.

Une note de préparation à l'IA sur laquelle vous pouvez agir. Qualité × classification × utilité — trois facteurs, une note. La note vous dit si vos données sont prêtes pour le déploiement de l'IA

avant de dépenser un euro en modèles ou en consultants. En dessous : les blocages spécifiques qui expliquent la note, chacun avec un niveau de priorité et une action de remédiation.

Signaux multi-sources qu'aucune vue en silo ne révèle. Le module Intelligence corrèle les conclusions entre les sources : « Vos exportations ERP contiennent des données d'identité dans les fichiers locaux sans organisation. » « La qualité de la base de données est excellente (A) mais la gouvernance échoue (F) — non-conformité invisible. » « L'désorganisation des fichiers rend la conformité structurelle impossible. » Ces signaux connectent des points que les outils source par source manquent.

Métriques de loi sur l'IA — pas de conformité à la loi sur l'IA. APOLLO ne prétend pas être conforme à la loi sur l'IA. Cela nécessite des systèmes de gestion des risques, une documentation technique, une surveillance humaine et des évaluations de conformité — tous basés sur le processus et en dehors du champ d'un scan de données. Ce qu'APOLLO fournit, ce sont les métriques de gouvernance des données de l'article 10 (complétude, cohérence, fraîcheur, contamination aux données d'identité) et les notes de posture cybersécurité de l'article 15 que les checkers de conformité ne peuvent pas produire. Les métriques que votre auditeur demandera — mesurées, pas déclarées.

Chaque note est transparente, reproductible et publiée. Pas de boîte noire.

LA COMPARAISON DES PRIX

	FRAMEWORKS DE PRÉPARATION À L'IA	CHECKERS DE CONFORMITÉ À LA LOI SUR L'IA	DSPM D'ENTREPRISE (MODULE IA)	APOLLO DATA AUDITOR
Coût annuel	Gratuit – 50 000 $ (conseil)	5 000 € – 100 000 €	100 000 $ – 250 000 €+	**2 999 € / an** (Starter)
Mesure la qualité des données pour l'IA	Non (maturité organisationnelle)	Non (déclaratif)	Partiel (inventaire de modèles)	**Oui — par source, par champ**
Identifie les blocages de l'IA	Non	Non	Partiel	**Oui — concrets, actionnables**
Métriques loi sur l'IA Art. 10	Non	Non (questionnaire)	Partiel	**Oui — automatisé à partir du scan**
Exposition à l'IA fantôme	Non	Non	Partiel (détection)	**Oui — cartographie PII des données exposées**

5. UN SCAN. UNE RÉPONSE. QUATRE DIMENSIONS.

La préparation à l'IA n'est pas une métrique isolée. Les données d'entraînement contaminées constituent également une violation de conformité (RGPD Art. 5, Art. 9). Un compte administrateur dormant avec accès à des ensembles de données prêts pour l'IA est à la fois une lacune de protection et un risque de qualité des données. Les fichiers obsolètes alimentant un pipeline d'IA constituent une exposition financière et une défaillance d'hygiène des données.

Les organisations qui traitent la préparation à l'IA, la conformité, la protection et le risque comme des flux de travail séparés se retrouvent avec des outils séparés, des budgets séparés et des lacunes de visibilité séparées. Le coût combiné dépasse 100 000 €/an — et les lacunes de visibilité demeurent.

APOLLO a été construit sur la prémisse que ces éléments constituent un seul problème, visible en un seul scan.

→ **Qualité & IA** — ce que ce document couvre. Note de préparation à l'IA, métriques de qualité des données, blocages de l'IA, radar de maturité des données, pré-conformité à la loi sur l'IA (Art. 10 + Art. 15), évaluation d'exposition à l'IA fantôme.

→ **Risque Privacy** — quantification financière en € et $, simulation d'impact de brèche, cartographie des données d'identité, combinaisons toxiques, zones à risque.

→ **Risque Conformité** — RGPD noté par article (Art. 5, 9, 30, 32), CCPA, NIS2, SOC2, DORA. Notes de A à F basées sur les données réelles. Plan de remédiation avec impact financier par action.

→ **Risque Protection** — résilience des sauvegardes, couverture de chiffrement, préparation aux ransomwares, hygiène du contrôle d'accès, identification des données ROT, simulation d'impact de brèche.

Aucun autre outil sous 5 000 €/an ne couvre tous les quatre. Le marché de l'IA a été construit autour de modèles et de plateformes. **APOLLO™ Data Auditor** existe parce que les données qui les sous-tendent déterminent s'ils réussissent ou échouent — et personne ne les mesure à un prix que les PME peuvent se permettre.

Sources : MIT — Nanda et al., « The GenAI Divide » (2025, via Fortune & Bloomberg) · Gartner AI Data Quality Analysis 2024–2025 · RAND Corporation AI Project Failures 2024 · IBM Cost of a Data Breach 2025 · TechShift Enterprise AI Readiness Report 2026 · EU AI Act (Regulation 2024/1689) · International AI Safety Report 2025 · Enzoic AD Password Security Report 2025

APOLLO™ Data Auditor

Tout fichier est un risque. Mesurez-le.

→ https://apollo.aiia-tech.com
→ GitHub : https://ggabrie2025.github.io/apollo_data_auditor/
→ contact@aiia-tech.com