Un modèle n'est pas souverain parce qu'il est ouvert

Trois chaînes de preuve qui décident si un système IA reste certifiable après son déploiement : poids inspectables, données documentables, versions gouvernables.

Points saillants

- "La souveraineté du modèle ne se lit pas dans le passeport de l'éditeur. Elle se lit dans trois chaînes de preuve instruites séparément : les poids que l'on peut inspecter et figer, les données d'entraînement que l'on peut documenter, les versions que l'on peut gouverner dans la durée. Open-weight ouvre la première porte. Il ne ferme ni la genèse, ni le cycle de vie."
- "Open-weight n'est pas open-source. La définition OSI Open Source AI 1.0 (28 octobre 2024) exige Data Information, Code et Parameters, avec une description suffisamment détaillée pour qu'un praticien qualifié recrée un système substantiellement équivalent. Llama 4, Mistral et DeepSeek publient leurs poids sans publier leurs données d'entraînement de manière reproductible. Ils sont open-weight, ils ne sont pas open-source au sens OSAID. OLMo 2 (Allen Institute for AI) l'est, et reste l'exemple frontier-adjacent le plus robuste."
- "L'AI Act et OSAID ne mesurent pas la même propriété. Le Training Data Summary Template publié par la Commission européenne le 24 juillet 2025 opérationnalise l'article 53(1)(d) du règlement IA : il exige une transparence narrative sur le contenu d'entraînement orientée vers l'opt-out copyright. L'OSAID exige une reproductibilité technique permettant la recréation d'un système équivalent. La conformité au premier ne produit pas la conformité au second."
- "Un modèle qu'on ne peut pas geler est un modèle qu'on ne peut pas certifier. La gouvernance des versions est l'axe le plus négligé de la souveraineté du modèle, et probablement le plus déterminant pour le déploiement régulé. Pour un dispositif logiciel médical certifié au titre du règlement IA et de MDR/IVDR intégrant une version donnée du foundation model, cette version exacte fait partie du périmètre certifié. Toute modification substantielle peut déclencher une révision de conformité."
- "La triade se convertit en matrice d'arbitrage CTO. Pour chaque plan (poids, données, versions), une question opérationnelle, une preuve minimale, un risque si absent. Cette matrice transforme la doctrine de constat en doctrine de

gouvernance, exactement comme la grille des trois verdicts l'a fait pour la couche matérielle dans le Volume 2."

- "La distinction qui tranche : open-weight est une catégorie commerciale, auditabilité composite est une catégorie d'ingénierie. Made in Europe modèle est une catégorie commerciale, souveraineté composite auditée du modèle est une catégorie d'ingénierie. Confondre l'une et l'autre dégrade la qualité des décisions architecturales prises sur cette base."

Mots clefs

- souveraineté du modèle open-weight OSAID 1.0
- GPAI Code of Practice AI Act article 53 Training Data Summary
- gouvernance des versions gel de version fork défensif
- OLMo 2 Mistral Apache 2.0
- matrice arbitrage CTO series: souverainete-numerique seriesIndex: 3 seriesTitle: "La souveraineté numérique" relatedTo:
- souverainete-architecturale souverainete-pile-pas-label
- eda-complement-essentiel-ia-agentique architecture-hexagonale-gouvernabilite
- gouvernance-ia-architecture

1. Introduction

Les deux premiers volumes de cette série ont posé une thèse en deux temps. Le Volume 1 a montré que la souveraineté de l'infrastructure d'exécution n'est pas un débat politique mais une condition de capitalisation de la performance, calculée par la garde et tracée dans le registre. Le Volume 2 est descendu dans la pile matérielle pour montrer que la souveraineté n'est pas un label mais une pile à sept strates, dont chacune appelle un verdict d'arbitrage autonome (acceptable, compensable, disqualifiante). Restait une couche dont les deux notes précédentes ont signalé l'importance sans la traiter : la couche modèle.

C'est l'objet de la présente note. Et la question doctrinale est plus piégée qu'elle n'y paraît, parce que la couche modèle est, parmi les sept strates, celle où le marketing produit le plus d'illusions de souveraineté. Mistral est français, donc Mistral est souverain. Llama 4 est open-weight, donc Llama 4 est libre. DeepSeek V4 est sous licence MIT, donc DeepSeek V4 est ouvert. Cohere et Aleph Alpha viennent de fusionner sous bannière

souveraine, donc l'Europe a son champion. Chacun de ces énoncés est, pris à la lettre, vérifiable. Et chacun est, pris comme conclusion, faux pour un déploiement régulé.

La thèse de cette note est sobre. *La souveraineté du modèle ne se lit pas dans le passeport de l'éditeur. Elle se lit dans trois chaînes de preuve : les poids que l'on peut inspecter et figer, les données d'entraînement que l'on peut documenter, les versions que l'on peut gouverner dans la durée. Open-weight ouvre la première porte. Il ne ferme ni la genèse, ni le cycle de vie.* En environnement régulé, cette différence n'est pas académique : elle décide si le système reste certifiable après son déploiement.

Domaine de validité. Comme dans les deux volumes précédents, la thèse vaut pour les systèmes d'IA déployés en environnement régulé européen (RGPD, HDS, NIS2, AI Act haut risque, MDR, IVDR), et pour lesquels la traçabilité des décisions, l'auditabilité par les organismes notifiés, et la stabilité juridique pluri-annuelle font partie des exigences fonctionnelles. Pour les usages non régulés, une partie des contraintes décrites ici se desserrent. Pour les usages régulés en santé clinique, en justice algorithmique, en infrastructure critique, et en services financiers à enjeu prudentiel, elles se renforcent.

Sous réserve de l'adoption formelle du texte Omnibus VII, dont l'accord politique provisoire a été annoncé le 7 mai 2026 et mentionné dans le Volume 2, le calendrier d'application des obligations haut risque est désormais à traiter comme mobile. Cela ne modifie ni l'article 53 GPAI, ni la structure du modèle de menace, ni le calendrier d'enforcement du GPAI Code of Practice à compter du 2 août 2026.

2. Trois confusions à dissiper

Avant d'instruire les trois chaînes de preuve, trois confusions doivent être nommées. Elles sont structurantes du faux débat actuel et elles produisent, dans les comités d'architecture, des conclusions inverses de la réalité réglementaire.

Première confusion : open-weight et open-source.

L'Open Source Initiative a publié le 28 octobre 2024 la version 1.0 de l'Open Source AI Definition. Le texte exige trois composants cumulatifs pour qu'un système d'IA soit qualifié d'open-source : *Data Information* suffisamment détaillée pour qu'un praticien qualifié recrée un système substantiellement équivalent à partir de données identiques ou similaires ; *Code* couvrant la totalité du pipeline (préparation des données, entraînement, validation, inférence) sous une licence approuvée par l'OSI ; et *Parameters* (les poids du modèle, et possiblement les checkpoints intermédiaires) sous des termes ouverts. Sur ces trois critères, la quasi-totalité des foundation models couramment qualifiés d'open-source dans la presse spécialisée n'est pas conforme. Llama 4 publie ses poids sous une licence non approuvée par l'OSI, contenant une clause de seuil de 700 millions d'utilisateurs actifs mensuels au-delà de laquelle une licence séparée doit

être négociée avec Meta, et des restrictions territoriales européennes sur les capacités multimodales. Mistral publie ses poids sous Apache 2.0, ce qui satisfait la composante *Parameters*, mais ne publie pas la composante *Data Information* dans une forme reproductible. DeepSeek publie ses poids sous licence MIT pleine, mais expose ses corpus d'entraînement par description plutôt que par accès. Aucun de ces trois éditeurs n'est OSAID-conforme. OLMo 2 (Allen Institute for AI) l'est, parce qu'il publie simultanément poids, code complet, données, recettes d'entraînement, logs et checkpoints. Quelques autres modèles s'en approchent (Pythia d'EleutherAI, listé par l'OSI parmi les références alignées sur la définition), mais OLMo 2 reste l'exemple frontier-adjacent le plus robuste pour un déployeur régulé en 2026.

Open-weight signifie que les poids sont téléchargeables. Open-source au sens OSAID signifie qu'un système substantiellement équivalent est reconstructible. Ce ne sont pas les mêmes objets. La valeur d'usage d'un modèle est dans ses poids ; la valeur d'audit d'un modèle est dans la trajectoire qui a produit ses poids. Pour un déploiement régulé, c'est la deuxième valeur qui détermine la défendabilité.

Deuxième confusion : conformité AI Act et conformité OSAID.

La Commission européenne a publié le 24 juillet 2025 le *Training Data Summary Template* (TDS Template), qui opérationnalise l'obligation de l'article 53(1)(d) du règlement IA. Cet article exige des fournisseurs de GPAI un summary suffisamment détaillé du contenu utilisé pour l'entraînement. Le template, dans sa forme actuelle, se contente d'une narration en haut niveau : taille des datasets, types de sources (web crawl, données licenciées, données sous licence ouverte), mesures de filtrage et de respect des opt-outs, catégories agrégées. Cette obligation est moins exigeante que celle d'OSAID 1.0, qui demande une description suffisamment précise pour permettre la recréation d'un système équivalent par un praticien qualifié. *Les deux régimes ne mesurent pas la même propriété.* La conformité Article 53 est une obligation de transparence narrative orientée vers le marché européen et l'exercice des droits d'opt-out par les ayants droit. La conformité OSAID est une obligation de reproductibilité technique orientée vers la communauté ouverte. La pratique récente confirme que cette asymétrie est exploitée : un fournisseur peut être conforme AI Act sans être conforme OSAID, et c'est aujourd'hui le cas du paysage entier des modèles ouverts hors quelques exceptions documentées par l'OSI dont OLMo 2 reste la plus robuste.

Cette asymétrie n'est pas un défaut de conception du règlement. Elle reflète un arbitrage politique explicite entre l'incitation à l'ouverture et la protection des intérêts commerciaux des éditeurs. Elle a une conséquence pratique pour le déployeur régulé : *la simple conformité Article 53 du fournisseur ne suffit pas à constituer une chaîne d'audit*. Le déployeur doit demander, contractuellement ou techniquement, des éléments de transparence supplémentaires. Ou bien il doit constater par construction que la chaîne d'audit n'est pas fermée et arbitrer en conséquence.

Troisième confusion : nationalité de l'éditeur et souveraineté de fabrique.

Le Volume 2 a montré que la nationalité du dernier assembleur ne résume pas la souveraineté de la chaîne. La même règle s'applique à la couche modèle, et avec une particulière sévérité. Un modèle développé en Europe par un éditeur européen, mais entraîné sur infrastructure dont le silicium est non européen, dont la mémoire haute bande passante est non européenne, et dont les datasets contiennent des fractions importantes de contenu produit hors UE, ne devient pas un modèle souverain par le seul fait de son éditeur. L'éditeur résout, à la couche modèle, une question de droit applicable, de gouvernance d'entreprise et de roadmap produit. Il ne résout pas la dépendance silicium ni la dépendance mémoire, qui sont structurellement non-européennes et qui ont fait l'objet du Volume 2.

Cette distinction est exactement homologue à celle qui a été faite dans le Volume 2 entre *souveraineté d'usage* et *souveraineté de fabrique* pour la couche cloud opérateur. Confondre les deux dégrade la qualité des décisions architecturales.

3. Première chaîne : auditabilité des poids

Auditer un modèle, ce n'est pas auditer ses poids. Cette formulation, qui inverse l'intuition courante, mérite un développement.

Les poids d'un modèle sont des objets numériques de très grande dimension, produits par un processus d'entraînement dont l'auditabilité dépend de la documentation de tout ce qui a contribué à les calculer. Les poids eux-mêmes ne portent aucune trace explicite des données qui les ont engendrés. Disposer des poids permet de faire de l'inférence, du fine-tuning, du red-teaming et de la mesure de propriétés opérationnelles a posteriori. Cela ne permet pas, à soi seul, de répondre aux questions que pose l'auditabilité au sens régulatoire : *quels biais documentés ce modèle hérite-t-il, quelles classes de contenu a-t-il vu pendant son entraînement, quelles mises à jour de sécurité ont été appliquées et quand ?*

L'auditabilité des poids, comme propriété structurée du système, exige donc une chaîne d'attestation qui descend en dessous des poids. Cette chaîne comporte au moins cinq composantes. *Premièrement*, l'identité cryptographique stable des poids (un hachage publié et signé par l'éditeur, dont l'absence transforme tout audit en argumentation par confiance). *Deuxièmement*, la traçabilité des checkpoints intermédiaires d'entraînement, qui permet de reconstituer la trajectoire d'apprentissage et de localiser l'apparition de propriétés émergentes. *Troisièmement*, la documentation des techniques d'alignement et de post-training, avec exposition des datasets d'alignement et des protocoles de red-teaming. *Quatrièmement*, la déclaration des modifications post-publication (correctifs de sécurité, ajustements de modération, fine-tunes officiels) avec

versioning sémantique strict. *Cinquièmement*, l'attestation matérielle de l'environnement d'inférence, qui rejoint le port de souveraineté du Volume 1 et la couche silicium du Volume 2.

Cette chaîne est aujourd'hui partiellement disponible chez certains éditeurs et entièrement opaque chez d'autres. OLMo 2 d'Allen AI publie poids, code d'entraînement, datasets, recettes, logs et checkpoints intermédiaires, ce qui en fait l'un des rares foundation models qui satisfait simultanément OSAID et un audit cryptographique de bout en bout. Mistral expose ses poids sous Apache 2.0 mais ne publie ni la totalité de ses checkpoints intermédiaires, ni l'intégralité de ses protocoles d'alignement. DeepSeek expose ses poids sous MIT et un rapport technique, mais sa chaîne d'alignement et l'historique de ses corrections post-publication ne sont pas pleinement exposés. Llama 4 expose ses poids sous une licence non-OSI avec des restrictions territoriales européennes et ne publie pas ses datasets d'alignement.

La distinction qui tranche : un modèle dont les poids sont téléchargeables est un modèle interrogeable, pas un modèle auditable. L'interrogation permet de mesurer le comportement observé en sortie ; l'audit permet de remonter aux conditions de production de ce comportement. Pour un dispositif logiciel médical, cette différence est précisément ce que l'AI Act haut risque exige et ce que les organismes notifiés codifient dans leurs grilles d'évaluation pour 2027.

4. Deuxième chaîne : souveraineté des données d'entraînement

Si l'auditabilité des poids fixe le périmètre de ce qu'on peut savoir sur un modèle, la souveraineté des données d'entraînement fixe le périmètre de ce qu'on peut affirmer sur sa licéité, sa représentativité et sa stabilité réglementaire.

Cette deuxième chaîne est celle où l'asymétrie entre les régimes réglementaires est la plus marquée. L'article 53(1)(d) du règlement IA, opérationnalisé par le TDS Template du 24 juillet 2025, demande aux fournisseurs de GPAI une description narrative des contenus utilisés pour l'entraînement. Le chapitre Copyright du GPAI Code of Practice du 10 juillet 2025 ajoute des engagements de respect des standards techniques d'opt-out (robots.txt et les futures spécifications IETF), de non-utilisation de sources notoirement problématiques, et de mécanismes de plaintes accessibles aux ayants droit. Ce régime est exigeant en termes de gouvernance copyright. Il ne constitue pas, à lui seul, une chaîne d'audit reproductible. Un fournisseur peut être conforme tout en n'exposant qu'une version résumée et agrégée de ses corpus d'entraînement.

L'OSAID 1.0 demande davantage : non pas une description résumée mais une description suffisamment détaillée pour qu'un praticien qualifié recrée un système

substantiellement équivalent. Dans la pratique de la communauté ouverte, cela se traduit par la publication des datasets sources eux-mêmes ou, à défaut, des manifestes précis (sources, dates de crawl, filtres de qualité, méthodes de déduplication, prompts de génération synthétique le cas échéant). Cette obligation de reproductibilité est plus exigeante que l'obligation de transparence du règlement IA. C'est précisément cette différence qui sépare un modèle interrogeable d'un modèle auditable.

Pour un déploiement en santé régulée, la distinction n'est pas académique. Ce qui est certifié au titre du règlement IA et de MDR/IVDR n'est pas le foundation model, c'est *un dispositif logiciel médical intégrant une version donnée d'un foundation model*. La validation clinique de ce dispositif exige, selon les orientations actuelles des organismes notifiés et les standards émergents pour l'articulation entre AI Act et MDR/IVDR, une caractérisation de la population d'entraînement (distribution démographique, géographique, clinique), une identification des sources de biais documentés, et une démonstration de la représentativité par rapport à la population cible. Cette caractérisation n'est pas réalisable à partir d'un summary AI Act ; elle exige les manifestes OSAID-grade. *Un modèle qui ne révèle pas ses données d'entraînement n'est pas auditable cliniquement, il est interrogeable en sortie*, et l'organisme notifié qui doit valider l'intégration du modèle dans le dispositif fait face à un trou de chaîne d'audit qu'aucune mesure de comportement en validation ne ferme.

L'objection prévisible est que la confidentialité commerciale et les contraintes de droit d'auteur empêchent la publication intégrale des corpus. C'est exact. La doctrine n'exige pas que tout modèle soit OSAID. Elle exige que *la décision de déployer un modèle non-OSAID dans un dispositif régulé soit prise en connaissance du delta d'auditabilité*, et que ce delta soit compensé par d'autres mécanismes : tests indépendants exhaustifs sur la population cible, contrats de transparence asymétrique avec le fournisseur, clauses de notification des modifications, ou trajectoire de migration vers un modèle plus auditable à mesure qu'il devient disponible. La compensation est l'aveu lucide d'une dépendance, pas son ignorance, exactement comme dans le verdict *compensable* du Volume 2.

Une seconde dimension de cette chaîne, structurellement importante, est la question des données synthétiques. Plusieurs des modèles récents incluent dans leur pipeline d'entraînement des datasets synthétiques produits par d'autres modèles, eux-mêmes parfois entraînés sur des corpus dont la chaîne d'audit n'est pas fermée. Cette récursivité multiplie les couches d'opacité. Auditer un modèle distillé exige d'auditer aussi le modèle teacher, et les conditions sous lesquelles le teacher a généré les données synthétiques. *La récursivité de l'entraînement multiplie les chaînes d'audit qu'il faut fermer simultanément*.

5. Troisième chaîne : gouvernance des versions

La troisième chaîne est la plus négligée dans le discours public, et probablement la plus déterminante pour le déploiement régulé. Un modèle n'est pas un objet figé. Il est un objet dont l'identité est toujours indexée par une version, et dont chaque version a un cycle de vie réglementaire propre.

L'article 53 du règlement IA et le chapitre Transparency du GPAI Code of Practice exigent une documentation conservée pendant au moins dix ans à compter de la mise sur le marché de chaque version d'un GPAI. Cette obligation a une conséquence souvent mal comprise : *un modèle déprécié reste juridiquement actif pendant une décennie après sa retraite commerciale*. L'éditeur peut cesser de le maintenir, mais il ne peut pas effacer ses obligations documentaires, et le déployeur qui l'a intégré dans un dispositif régulé reste lui-même tenu de pouvoir présenter, sur dix ans, l'audit du modèle dans la version qui était utilisée au moment de chaque décision automatisée historique.

Cette propriété temporelle interagit avec une réalité industrielle bien documentée : les éditeurs de foundation models déprécient leurs modèles à un rythme rapide. DeepSeek a annoncé la retraite des endpoints API deepseek-chat et deepseek-reasoner pour le 24 juillet 2026, soit dix-huit mois après la mise en service de la génération qu'ils servent. Mistral renouvelle sa gamme à un rythme de plusieurs versions majeures par an. Meta a lancé Muse Spark en avril 2026, signe que la trajectoire Llama n'est pas une ligne stable mais une stratégie produit révisable. *Le rythme de dépréciation des modèles est structurellement plus court que la durée d'archivage réglementaire qu'ils déclenchent.* Pour un dispositif logiciel médical exploité dix ans, le foundation model qui en fait partie aura été déprécié sept à huit fois.

Cette désynchronisation produit un énoncé doctrinal qui doit être tenu pour ce qu'il est, c'est-à-dire un critère de certifiabilité avant même d'être un critère de gouvernance.

Un modèle qu'on ne peut pas geler est un modèle qu'on ne peut pas certifier.

Si la version exacte des poids utilisée au moment de la conformité initiale ne peut pas être archivée, reconstruite ou rejouée à l'identique, alors la conformité elle-même perd son objet. La certification porte sur un dispositif logiciel médical intégrant une version donnée du foundation model. Sans capacité de gel, le système certifié et le système opérationnel divergent à chaque mise à jour silencieuse de l'éditeur, et la conformité affichée devient une fiction documentaire.

Trois conséquences architecturales en découlent.

1. *Premièrement*, le **gel de version pour conformité**. Pour un dispositif logiciel médical qui obtient un marquage CE au titre du règlement IA et de MDR/IVDR, la version exacte du foundation model intégrée au moment de la conformité initiale fait partie du périmètre certifié. Toute modification substantielle (changement de version, fine-tune supplémentaire, mise à jour de sécurité majeure) déclenche potentiellement une révision de conformité. Le déployeur ne peut pas simplement suivre la roadmap du fournisseur ; il doit décider, version par version, si la mise à jour est appliquée, dans quel délai, et selon quel protocole de re-validation. Cette décision est une primitive d'architecture exposée comme port externe au système, dans la lignée de la doctrine hexagonale développée précédemment dans cette série.
2. *Deuxièmement*, le **fork défensif**. Lorsque le fournisseur déprécie un modèle dont une version gelée est encore en exploitation chez des déployeurs régulés, deux scénarios coexistent. Le fournisseur maintient les anciennes versions disponibles via API ou téléchargement, ou bien il les retire. Pour les modèles open-weight, le fork défensif est techniquement praticable : le déployeur télécharge la version exacte des poids, l'archive dans son propre périmètre, et continue à l'exploiter localement après dépréciation par l'éditeur. Pour les modèles purement API (typiquement les frontières propriétaires), ce fork n'est pas possible et la conformité dépend de la politique de rétention du fournisseur, contractuellement ou pas. *La capacité de fork défensif transforme un modèle externe en composant interne archivable*.
3. *Troisièmement*, la **chaîne de migration auditée**. Quand un déployeur décide de basculer d'une version à la suivante, la transition n'est pas neutre. Elle exige une démonstration de conformité de la nouvelle version sur les exigences fonctionnelles du système, une comparaison de performances sur la population cible, une analyse des biais hérités ou modifiés, et selon le contexte régulatoire, une nouvelle évaluation par l'organisme notifié. Cette chaîne de migration est une discipline d'ingénierie en soi, distincte de la mise à jour applicative classique. Elle repose architecturalement sur la capacité du système à maintenir simultanément deux versions actives pendant la phase de transition, à comparer leurs sorties sur des cas représentatifs, et à basculer de manière atomique avec point de rollback documenté.

La distinction qui tranche : *la version d'un modèle n'est pas un détail, c'est l'objet réglementaire*. La doctrine de la *gouvernance-as-architecture* développée précédemment dans cette série posait que les exigences de gouvernance ne sont pas des couches documentaires séparables du système ; elle se prolonge ici dans la couche modèle avec une exigence supplémentaire : la version est l'unité de discours réglementaire, et l'architecture doit la rendre explicitement manipulable.

6. Matrice d'arbitrage CTO

Une doctrine utile ne peut pas rester descriptive. Elle doit produire un instrument de décision tenable en comité d'architecture. La triade des trois chaînes se convertit en une matrice à quatre colonnes : le plan de souveraineté concerné, la question opérationnelle que le CTO doit pouvoir formuler, la preuve minimale à exiger du fournisseur ou à produire par construction, et le risque qui se matérialise si la preuve manque.

Chaîne	Question CTO	Preuve minimale	Risque si absent
Poids	Puis-je inspecter et figer le modèle effectif ?	Checksum signé, licence d'archive locale, artefact téléchargeable, fine-tune auditable	Dépendance opaque
Données	Puis-je défendre la genèse du modèle ?	TDS enrichi, Data Information OSAID-grade, audit indépendant des sources, documentation des biais	Biais non assignable
Versions	Puis-je maintenir la conformité dans la durée ?	Gel contractuel, préavis EOL, protocole de migration validé, fork défensif	Rupture de certification

Cette matrice est l'instrument concret de la triade. Elle permet, pour chaque foundation model envisagé, de produire une fiche d'arbitrage en trois lignes que le comité d'architecture peut signer ou refuser. Quand la preuve minimale est obtenue ou produisible, la chaîne est verte. Quand elle ne l'est pas, le risque correspondant doit être nommé, accepté formellement par le décideur, ou compensé par un mécanisme alternatif explicite. Quand elle ne peut être ni obtenue ni compensée, le modèle ne franchit pas la porte du système régulé.

La matrice ne sélectionne pas un fournisseur ; elle force la traçabilité de la décision. C'est sa valeur architecturale propre, et c'est exactement le rôle que la grille des trois verdicts (acceptable, compensable, disqualifiante) joue pour la couche matérielle du Volume 2.

7. Trois cas de test sur la matrice

La matrice ne se valide pas en panorama marché. Elle se valide sur des cas de test concrets. Trois trajectoires actuellement disponibles pour un déployeur régulé européen en 2026 servent ici d'instruction de la matrice, pas de classement de fournisseurs.

Mistral est l'éditeur européen de référence pour les foundation models open-weight. La gamme actuelle (Large 3 publié en décembre 2025 sous Apache 2.0, Small 4 publié en mars 2026 sous Apache 2.0) satisfait la chaîne *Poids* par la disponibilité publique des artefacts et par une licence permettant le fork défensif. Sur la chaîne *Données*, Mistral devra publier, comme tout fournisseur de GPAI opérant sur le marché européen, les informations requises par l'article 53 du règlement IA et le TDS Template ; cela ne signifie pas publication OSAID-grade des données d'entraînement. Sur la chaîne *Versions*, l'exposition via Hugging Face rend le gel contractuel praticable, mais le rythme de renouvellement exige du déployeur un protocole de migration auditée explicite. Le compute d'entraînement repose sur des accélérateurs NVIDIA, ce qui place Mistral dans la situation de souveraineté composite du Volume 2 : éditeur sous droit européen, dépendance silicium non-européenne en transition. *Sur la matrice, Mistral est compensable*.

OLMo 2 d'Allen Institute for AI est l'exception structurelle. Le modèle satisfait pleinement les trois chaînes : poids publics sous Apache 2.0, code d'entraînement complet, datasets sources publiés, recettes, logs et checkpoints intermédiaires accessibles. C'est aujourd'hui le foundation model frontier-adjacent le plus robuste pour qualifier comme OSAID-conforme par construction. Le compromis est réel : OLMo 2 reste en dessous de la frontière de performance des modèles propriétaires de plus grande échelle pour les tâches les plus exigeantes, et son éditeur est américain, ce qui replace la question de souveraineté juridique dans le périmètre du Volume 1. *Sur la matrice, OLMo 2 est acceptable* pour les déploiements dont la performance frontier n'est pas le critère discriminant.

Cohere et Aleph Alpha, depuis l'annonce du 24 avril 2026 de leur fusion, constituent une entité transatlantique dont la pile d'exécution prévue passe par STACKIT, le cloud souverain de Schwarz Digits. La gamme Cohere n'est pas open-weight ; les modèles sont accessibles par API. Sur la matrice : chaîne *Poids* non satisfaite par téléchargement (fork défensif non praticable hors contrat ad hoc) ; chaîne *Données* non exposée au standard OSAID ; chaîne *Versions* sous contrôle de l'éditeur fusionné, partiellement compensée par la garantie d'exécution sur infrastructure souveraine européenne. *Sur la matrice, ce profil est compensable sur l'infrastructure, non couvert sur l'auditabilité du modèle sans contrat dédié*.

Ces trois cas de test illustrent le constat doctrinal central : *aucun ne ferme la triade sans concession*. L'instrument n'élit pas un fournisseur, il instruit la décision.

8. Trois verdicts pour la couche modèle

Le Volume 2 a posé une grille à trois verdicts pour la couche matérielle. Cette grille s'applique à la couche modèle avec une instanciation propre.

Acceptable. La concession sur les chaînes existe, mais elle est techniquement substituable à coût et délai raisonnables, et elle ne crée pas d'asymétrie de pouvoir que le fournisseur peut exploiter unilatéralement. La portabilité vers un autre modèle est documentée et bornée. La dépendance existe, mais elle est symétrique au sens où la rupture serait coûteuse pour les deux parties.

Compensable. La concession n'est pas substituable à court terme, mais elle peut être couverte par des mécanismes contractuels ou opérationnels vérifiables. La compensation prend la forme de quatre dispositifs cumulatifs : le contrat-cadre exposant les engagements du fournisseur au-delà du minimum AI Act ; le fork défensif systématique (poids archivés en local pour chaque version déployée en production régulée) ; le protocole de re-validation déclenché par toute modification substantielle ; et la trajectoire de migration documentée vers un modèle plus auditable à mesure que l'écosystème mûrit. La compensation est l'aveu lucide d'une dépendance, pas son ignorance.

Disqualifiante. La concession peut interrompre, altérer ou rendre non défendable un système critique sans recours réaliste. Exemple typique : un modèle dont le fournisseur peut, par mise à jour silencieuse, modifier le comportement de la version déployée sans notification, sans possibilité de rollback contractuel, et sans manifeste public des modifications. Pour un dispositif logiciel médical, cette propriété rend la conformité MDR/IVDR impossible à maintenir dans le temps. Autre exemple : un modèle dont l'architecture d'attestation matérielle est conditionnée par une autorité externe au déployeur, sans clause d'isolation contre l'évolution réglementaire de la juridiction émettrice, et qui rejoint alors la disqualification décrite à la couche silicium dans le Volume 2. *Si la concession est disqualifiante, l'architecture doit être revue, pas adoucie par un discours*.

9. Pourquoi le paradigme « Made in Europe modèle » est insuffisant

Le débat européen sur la souveraineté des foundation models glisse vers une logique binaire homologue à celle qui a été critiquée dans le Volume 2 pour la couche matérielle : modèle européen ou modèle non européen. Cette grille est insuffisante pour deux raisons symétriques.

Un modèle développé en Europe par un éditeur européen, mais entraîné sur infrastructure dont le silicium est non européen, dont la mémoire haute bande passante est non européenne, et dont les datasets contiennent des fractions importantes de contenu produit hors UE, ne devient pas un modèle souverain par le seul fait de son éditeur. La nationalité du dernier intégrateur ne résume pas la souveraineté de la chaîne de fabrique du modèle. Acheter un modèle européen ne fabrique pas un modèle européen au sens des trois chaînes.

Inversement, l'usage d'un modèle non européen dans une architecture qui dispose d'une gouvernance locale forte (fork défensif, archive locale des poids, protocole de re-validation, contrat de transparence supplémentaire), d'une chaîne d'audit fermée par l'aval (tests indépendants exhaustifs sur la population cible, monitoring de conformité runtime), et d'une trajectoire de migration vers un modèle plus auditable, peut être plus défendable qu'un modèle européen opaque dont aucune des trois chaînes n'est instruite.

La distinction qui tranche : *open-weight est une catégorie commerciale, auditabilité composite est une catégorie d'ingénierie. Made in Europe modèle est une catégorie commerciale, souveraineté composite auditée du modèle est une catégorie d'ingénierie.* Ce ne sont pas les mêmes objets, et les confondre dégrade la qualité des décisions architecturales.

Cette homologie avec le Volume 2 n'est pas accidentelle. Elle traduit le fait que le pattern doctrinal de la souveraineté composite, posé pour la couche matérielle, est strictement transposable à la couche modèle parce qu'il décrit la même structure de problème : une stratification de dépendances dont chacune appelle un verdict d'arbitrage propre, et dont la composition produit, ou ne produit pas, une auditabilité défendable.

10. Conclusion

Le faux dilemme « modèle européen contre modèle américain » est l'analogue, pour la couche modèle, du faux dilemme « performance contre souveraineté » que le Volume 1 a déconstruit pour la couche cloud. Les deux suppositions partagent le même vice : elles traitent la souveraineté comme un attribut binaire d'un objet, alors qu'elle est une propriété structurée d'une architecture.

La souveraineté du modèle ne se lit pas dans le passeport de l'éditeur. Elle se lit dans trois chaînes de preuve : les poids que l'on peut inspecter et figer, les données d'entraînement que l'on peut documenter, les versions que l'on peut gouverner dans la durée. Open-weight ouvre la première porte. Il ne ferme ni la genèse, ni le cycle de vie. En environnement régulé, cette différence n'est pas académique : elle décide si le système reste certifiable après son déploiement.

La triade qui structure cette note se convertit en matrice d'arbitrage CTO : trois questions, trois preuves minimales, trois risques. Ce n'est pas une grille de notation, c'est une

discipline de traçabilité de la décision qui transforme l'évaluation d'un foundation model en une instruction systématique des trois chaînes, suivie d'un verdict (acceptable, compensable, disqualifiante) propre à chaque déploiement. Cette discipline est exigeante, mais elle est la seule techniquement défendable pour construire des systèmes IA régulés sans tomber ni dans l'illusion de l'open-weight comme garantie de souveraineté, ni dans le rejet idéologique des modèles dont la chaîne de fabrique n'est pas intégralement européenne.

Le paysage européen 2026 confirme cette analyse. Mistral consolide sa trajectoire frontier indépendante en restant dépendant du silicium et de la mémoire non européens. La fusion Cohere-Aleph Alpha trace une voie de souveraineté composite transatlantique adossée au cloud STACKIT, en renonçant explicitement à la compétition frontier en isolation. OLMo 2 démontre qu'une auditabilité OSAID complète est techniquement réalisable, au prix d'une concession sur la performance frontier. Aucune de ces trajectoires ne ferme la triade pour tous les déploiements. *Aucune ne le pourra*, parce que la souveraineté du modèle, comme celle de l'infrastructure et celle de la pile matérielle, est une propriété composite qui se construit, par déploiement, et non un état pur qui se choisirait.

Une certification SecNumCloud immunise l'opérateur, pas le silicium qu'il opère ; le silicium opère un modèle, pas la chaîne d'auditabilité qui le rend défendable ; et un modèle qu'on ne peut pas geler est un modèle qu'on ne peut pas certifier. L'extension de la doctrine de la souveraineté architecturale à la couche modèle ne change pas la nature du critère, elle l'instancie sur une nouvelle strate de la pile.

Le Volume 4 de cette série traitera la couche énergétique, la dernière des quatre couches identifiées dans le Volume 1 et signalée comme limite authentique de la doctrine. La séquence sera alors complète, et l'arbitrage architectural dont la triade est l'instrument pourra être instruit en pleine connaissance de cause sur l'ensemble de la pile : infrastructure, fabrique, modèle, et énergie qui les alimente toutes.

Sources et références

1. **Open Source Initiative** : *The Open Source AI Definition 1.0*, 28 octobre 2024. Définit les composants requis pour qu'un système d'IA soit qualifié d'open-source : Data Information, Code, Parameters. opensource.org/ai/open-source-ai-definition

2. **AI Office, Commission européenne** : *General-Purpose AI Code of Practice*, version finale publiée le 10 juillet 2025. Trois chapitres : Transparency, Copyright, Safety and Security. Application 2 août 2025 (nouveaux modèles), enforcement 2 août 2026, modèles antérieurs au 2 août 2025 ont jusqu'au 2 août 2027 pour se mettre en conformité. digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

3. **Commission européenne** : *Training Data Summary Template*, publié le 24 juillet 2025, opérationnalisant l'article 53(1)(d) du règlement IA pour les fournisseurs de GPAI. Demande une description narrative haut niveau du contenu d'entraînement orientée vers l'opt-out copyright.

4. **Conseil de l'Union européenne** : *Artificial Intelligence: Council and Parliament agree to simplify and streamline rules*, communiqué du 7 mai 2026. Accord politique provisoire sur le paquet Omnibus VII, sous réserve d'adoption formelle du texte final. Le calendrier est interprété par certaines sources juridiques comme un report des obligations Annexe III au 2 décembre 2027 et Annexe I au 2 août 2028 ; la rédaction définitive reste à surveiller. consilium.europa.eu

5. **Mistral AI** : page produit Mistral Large 3, mis à disposition en décembre 2025 sous licence Apache 2.0, architecture Mixture-of-Experts à 675 milliards de paramètres totaux et 41 milliards actifs par token, entraîné sur GPU NVIDIA H200. mistral.ai

6. **Mistral AI** : page produit Mistral Small 4, publié en mars 2026 sous licence Apache 2.0, architecture Mixture-of-Experts à 119 milliards de paramètres totaux et 6 milliards actifs par token. mistral.ai

7. **Meta AI** : Llama 4 publié le 5 avril 2025 sous Llama 4 Community License, non approuvée par l'Open Source Initiative. Clause de seuil de 700 millions d'utilisateurs actifs mensuels. Restrictions territoriales européennes sur les capacités multimodales. ai.meta.com/blog/llama-4-multimodal-intelligence/

8. **DeepSeek** : annonce officielle de la retraite des endpoints API deepseek-chat et deepseek-reasoner pour le 24 juillet 2026, exemple concret de cycle de vie commercial d'un foundation model utilisé en production.

9. **Allen Institute for AI** : famille OLMo 2, identifiée par OSI comme conforme OSAID 1.0. Poids, code d'entraînement, datasets, recettes, logs et checkpoints publiés sous Apache 2.0. allenai.org

10. **EleutherAI** : famille Pythia, également identifiée par OSI comme conforme OSAID 1.0, instruments de référence pour la recherche en interprétabilité et la reproductibilité des trajectoires d'entraînement.
11. **Cohere et Aleph Alpha** : annonce du 24 avril 2026 de la fusion entre Cohere (Toronto) et Aleph Alpha (Heidelberg), avec exécution sur le cloud souverain STACKIT opéré par Schwarz Digits. Pivot stratégique d'Aleph Alpha hors de la compétition frontier model documenté en 2024.
12. **Latham & Watkins** : *EU AI Act: GPAI Model Obligations in Force and Final GPAI Code of Practice in Place*, 30 juillet 2025. Analyse du régime de fines (jusqu'à 15 millions d'euros ou 3 % du chiffre d'affaires global, article 101) applicable aux non-conformités GPAI à compter du 2 août 2026. lw.com