

souveraineté et ce qu'ils coûtent

Héberger en France ne protège pas du CLOUD Act. Les trois niveaux de souveraineté IA, le scénario adapté à ta taille et son coût mensuel réel, chiffré.

BLYX-AGENCY.COM · 27 JUILLET 2026 · PAR MELVIN BRETON

Le CLOUD Act permet à la justice américaine d'exiger tes données à tout éditeur de droit américain, où que soient ses serveurs. « Souverain » recouvre trois niveaux : la résidence des données, qui ne protège de rien, la juridiction européenne de l'éditeur, environ 20 euros par personne et par mois, et le contrôle des poids sur ta machine, dès environ 180 euros par mois. Ce guide chiffre chaque niveau ; [le versant équipement du même sujet](#) complète la décision.

1. Ce que le CLOUD Act change pour tes données

Le CLOUD Act est une loi fédérale américaine de 2018. Elle oblige tout fournisseur soumis au droit américain à remettre les données qu'il contrôle, quel que soit le pays où se trouve le serveur. La juridiction suit le fournisseur, pas le datacenter.

Concrètement, chaque devis, contrat ou fichier client collé dans un outil américain part chez une société de droit américain. Et cette société reste saisissable par la justice américaine, même quand ses serveurs tournent à Paris.



La démonstration a eu lieu devant le Sénat français. Le 10 juin 2025, le

Blyx juridique de Microsoft France, interrogé sous serment, a  reconnu ne pas pouvoir garantir que les données de clients français ne seraient jamais transmises aux autorités américaines. Microsoft héberge pourtant en France. L'adresse du serveur ne protège pas.

La nuance qui évite de dramatiser : il faut un mandat validé par un juge américain, sur des motifs sérieux. Ce n'est pas une aspiration permanente de tes fichiers. C'est un risque juridique structurel, que ni ton contrat ni la localisation du serveur ne neutralisent.

Ce qui déclenche le CLOUD Act, ce n'est pas l'adresse du serveur. C'est la nationalité juridique de l'entreprise qui l'opère.

2. Souverain veut dire trois choses différentes

Le mot est utilisé pour vendre trois niveaux de protection très inégaux. Savoir les distinguer, c'est déjà éviter 80 % des mauvaises décisions.

Niveau	Ce que ça garantit	Ce que ça ne garantit pas	Exemple type
1. Résidence des données	Les fichiers sont stockés physiquement en Europe	Rien face au CLOUD Act si l'éditeur est américain	Une option « EU data residency » chez un éditeur américain





2. Juridiction de l'éditeur	L'entreprise relève du droit européen, donc du RGPD seul	Que le modèle tourne chez toi, tu dépends encore d'un tiers	Mistral hébergé en France
3. Contrôle des poids	Le modèle tourne sur ta machine, sans appel sortant	Rien de plus, c'est le niveau maximal	Un modèle ouvert installé sur ton serveur

Conséquence pratique : passer du niveau 1 au niveau 2 change tout sur le plan juridique, et ne coûte quasiment rien. C'est le meilleur rapport effort sur protection du marché, et c'est là que 90 % des entreprises devraient commencer.

3. Commencer par le niveau deux

Le niveau 2 couvre l'écrasante majorité des besoins réels : rédaction, synthèse, analyse de documents, assistance quotidienne. Basculer ces usages vers un éditeur européen se fait en une journée, sans projet technique et sans recrutement.

Côté prix, Mistral Le Chat Team est facturé 19,99 dollars par utilisateur et par mois en engagement annuel, 24,99 dollars en mensuel (vérifié le 30/08/2026). C'est l'ordre de grandeur d'un abonnement américain équivalent. Pour une équipe de 15 personnes, on parle d'environ 300 euros par mois.

Avant de signer, lis deux clauses : celle qui dit si tes échanges entraînent le modèle, et celle qui fixe la durée de conservation. C'e



4. Les trois scénarios selon ta taille

Voici la grille. Les coûts sont des ordres de grandeur constatés à l'été 2026, hors accompagnement, vérifiés le 30/08/2026.

	PME (moins de 50 pers.)	ETI (50 à 500 pers.)	Grand groupe (500 et plus)
Solution	Abonnement éditeur européen	Modèle ouvert sur GPU loué en France	Infrastructure dédiée
Exemple	Mistral Le Chat Team	Mistral Small, Llama, Qwen	Cluster interne ou colocation
Où tournent les données	Cloud français	Serveur loué, France	Ton datacenter
Coût mensuel	~20 € par personne	~600 à 2 400 € (GPU seul)	100 k€ et plus en investissement
Délai de mise en route	Une journée	1 à 3 semaines	6 à 18 mois
Compétence requise	Aucune	Un profil technique	Équipe dédiée
Niveau de souveraineté	2	3	3

Le coût du scénario ETI mérite une précision, parce que deux chiffres circulent sur ce site et les deux sont justes. La ligne du tableau couvre



la location du GPU seul, pour un modèle de taille moyenne opéré par

technique. Un déploiement de production sur deux nœuds,  Blyx 

avec supervision et astreinte, revient plutôt entre 18 000 et 28 000

euros par mois : c'est l'ordre de grandeur retenu dans le versant

équipement. Entre les deux, tout dépend de la taille du modèle et du

niveau de service exigé.

Côté grands groupes, même eux commencent rarement par le cluster

interne. Le ministère des Armées a notifié le 16 décembre 2025 un

accord-cadre avec Mistral, déployé sur infrastructure contrôlée par la

France. L'État fait tourner son assistant Albert sur des modèles ouverts

hébergés en France. Ce sont des modèles de taille raisonnable sur

infrastructure maîtrisée, pas des monstres à 2 800 milliards de

paramètres.

5. Le coût réel d'un GPU loué en France

Dans ce scénario, le modèle tourne sur une machine que tu loues en

France. Personne d'autre n'y accède, et aucune donnée ne sort de

l'instance. Le coût dépend entièrement du GPU choisi, donc de la taille

du modèle. Prix constatés, vérifiés le 30/08/2026 :

Configuration	Prix horaire constaté	Coût en 24/7	Coût en heures ouvrées
L4 (24 Go)	~0,85 €/h	~620 €/mois	~180 €/mois
A100 (80 Go)	1,52 à 1,85 €/h	~1 100 à 1 350 €/mois	~320 à 390 €/mois
H100 (80 Go)	2,73 à 3,30 €/h	~2 000 à 2 400 €/mois	~580 à 700 €/mois



éteint la nuit et le week-end coûte trois fois moins. Beaucoup de projets sont abandonnés sur le calcul en 24/7 alors que l'usage réel va de 9 h à 19 h.

Ce scénario s'impose quand tu manipules des données réglementées, du secret industriel, ou une obligation contractuelle de non-transfert. Côté volume, l'auto-hébergement d'un petit modèle sur un GPU unique devient rentable au-delà d'environ 5 millions de tokens par mois, comparé aux mêmes appels facturés à l'usage (vérifié le 30/08/2026). Pour un gros modèle réparti sur plusieurs nœuds, le seuil monte vers 800 millions de tokens par mois, l'ordre de grandeur retenu dans le versant équipement.

OVHcloud, Scaleway et Outscale louent du GPU en France. Le comparatif détaillé des hébergeurs, qualification SecNumCloud comprise, est traité dans le versant équipement : inutile de le dupliquer ici.

6. Quel modèle pour quelle machine

La règle de dimensionnement est simple : un modèle quantifié en 4 bits occupe environ la moitié de son nombre de paramètres en gigaoctets de VRAM, plus une marge pour le contexte.



3 à 8 milliards	~8 Go	L4	Extraction, classement, résumé court
13 à 30 milliards	16 à 24 Go	L4, L40S	Rédaction, analyse de documents, support
70 milliards et plus	48 Go et plus	A100, H100	Raisonnement complexe, code
2 800 milliards (Kimi K3)	~1 400 Go	64 GPU et plus	Hors de portée sans datacenter dédié

Sur le français, les modèles Mistral restent la référence : ils sont nativement entraînés sur un corpus français de qualité, là où les modèles asiatiques ou américains traitent le français comme une langue secondaire.

Le réflexe qui fait gagner le plus de temps : commencer par le plus petit modèle qui passe le test, pas par le plus gros que le budget autorise. Un modèle de 8 milliards de paramètres bien intégré bat un modèle de 70 milliards mal branché.

7. Les cinq pièges qui coûtent cher

1. **Utiliser le chat gratuit d'un éditeur pour du travail sérieux.** Les versions gratuites s'entraînent presque toujours sur ce que tu éc



Le modèle téléchargé et le chat en ligne du même éditeur sont deux **réalités opposées**. Le détail des clauses à vérifier est dans [l'article sur les risques de ChatGPT en entreprise](#).

2. **Confondre hébergement et juridiction.** Le piège numéro un, démonté en section 1 : l'adresse du serveur ne te protège de rien.
3. **Croire qu'un modèle ouvert est gratuit.** Les poids le sont. Le GPU, l'exploitation, la supervision et les mises à jour ne le sont pas.
4. **Dimensionner en 24/7 par défaut.** C'est ce qui fait capoter la plupart des dossiers, alors que l'usage réel tient sur les heures ouvrées.
5. **Croire que le modèle fait le produit.** Le modèle n'est que le moteur. La valeur vient de son branchement sur tes vrais systèmes : CRM, facturation, logiciel métier. C'est là que se joue l'écart entre un gadget et un gain mesurable.

8. Par où commencer cette semaine

Trois actions, selon où tu en es.

1. **Faire l'inventaire.** Liste les outils IA réellement utilisés dans l'entreprise, y compris ceux que personne n'a validés. C'est presque toujours plus large que prévu, et le versant équipement chiffre ce que cet usage non encadré coûte.
2. **Traiter le niveau 2 en premier.** Basculer les usages bureautiques vers un éditeur européen se fait en une journée et supprime l'exposition au CLOUD Act sur 90 % du volume.
3. **Isoler ce qui justifie vraiment le niveau 3.** En général, un ou deux processus : documents contractuels, données de santé, R&D. C



sur ce périmètre restreint qu'un modèle auto-hébergé se justifie,



Avant de brancher quoi que ce soit sur des données sensibles, passe chaque connexion au crible de la [grille d'arbitrage IA et données juridiques](#). Garde aussi en tête que les obligations de transparence de l'AI Act s'appliquent depuis le 2 août 2026 : informer les personnes qu'elles s'adressent à une IA, et marquer les contenus générés destinés à être publiés.

Ce tri entre un niveau 2 généralisé et un niveau 3 ciblé est exactement le travail d'un [audit IA](#) : cartographier les usages, classer les données par sensibilité, puis chiffrer les deux ou trois processus qui justifient un modèle chez toi.

La question n'est pas « quel est le meilleur modèle ». Elle est « quelles données ne doivent jamais sortir, et quel est le chemin le plus court pour les protéger ».

Questions fréquentes

Un hébergement en France suffit-il à protéger du CLOUD Act

Non. La loi suit la nationalité juridique du fournisseur, pas l'adresse de ses serveurs. Un éditeur américain qui héberge à Paris reste saisissable par la justice américaine, ce que Microsoft France a confirmé sous serment devant le Sénat en juin 2025. La protection commence quand l'éditeur relève du droit européen.



La justice américaine, sur mandat validé par un juge et pour des motifs sérieux, auprès de tout fournisseur soumis au droit américain. Ce n'est pas une surveillance continue de tes fichiers. C'est un risque juridique structurel, que ni ton contrat ni la localisation du serveur ne neutralisent.

Combien coûte une IA hors de portée du CLOUD Act

Environ 20 euros par personne et par mois pour un abonnement chez un éditeur européen, le niveau qui suffit à la plupart des TPE et PME. Pour un modèle qui tourne sur une machine louée en France, compte environ 180 euros par mois pour un GPU d'entrée de gamme utilisé aux heures ouvrées (vérifié le 30/08/2026).

À partir de quel volume l'auto-hébergement devient-il rentable

Autour de 5 millions de tokens par mois pour un petit modèle sur un GPU unique, comparé aux mêmes appels facturés à l'usage. Pour un gros modèle réparti sur plusieurs nœuds, le seuil monte vers 800 millions de tokens par mois. En dessous, l'abonnement ou l'API restent la bonne réponse.

