

Guide décideur : l'IA souveraine en entreprise

Ce que change Kimi K3 pour héberger l'IA en interne — 3 scénarios chiffrés, comparatif des clouds souverains français, checklist RGPD et AI Act.

BLYX-AGENCY.COM · 19 JUILLET 2026



GUIDE DÉCIDEUR : l'IA souveraine en entreprise — par Pierre Cordelier.

Ce que change Kimi K3, et comment décider sans se tromper.

1. Ce qui vient de se passer (et pourquoi ça te concerne)

Le 16 juillet 2026, Moonshot AI a publié **Kimi K3** : le plus gros modèle open source jamais rendu public.

Les chiffres qui comptent :

- ✓ **2 800 milliards de paramètres** (architecture MoE, ~50 milliards actifs par requête)
- ✓ **1 million de tokens** de contexte — soit environ 1 500 pages d'un coup
- ✓ **Multimodal** : texte, images et vidéo en entrée
- ✓ **Poids téléchargeables le 27 juillet 2026**



Le point important n'est pas la taille. C'est le **niveau**.



Sur l'Intelligence Index d'Artificial Analysis — un classement



indépendant, pas un benchmark maison — Kimi K3 obtient **57**. Claude Opus 4.8 est à ~56, GPT-5.5 à ~57. Claude Fable 5 reste devant à ~60, GPT-5.6 Sol à 58,9.

Traduction : **pour la première fois, un modèle que tu peux télécharger et faire tourner chez toi joue dans la même division que les meilleurs modèles fermés**. Il est même n°1 mondial sur la génération de code front-end (LMArena, ~76 % de victoires).

***Attention** : la plupart des autres benchmarks sont auto-déclarés par Moonshot. Les évaluations indépendantes complètes arrivent après la sortie des poids. Ne signe pas un chèque sur la foi d'un communiqué de presse.*

2. Le vrai sujet : la juridiction, pas le vol

Il faut être précis, parce que le débat est pollué par beaucoup d'approximations.

Ce qui est FAUX : « Anthropic et OpenAI s'entraînent sur tes données. »

Sur les API entreprise, c'est contractuellement faux. OpenAI ne s'entraîne pas sur les données API depuis mars 2023. Anthropic ne conserve pas les échanges par défaut au-delà d'environ 30 jours et ne les utilise pas pour l'entraînement sans autorisation. Les deux proposent du Zero Data Retention sur demande.



Ce qui est VRAI, et beaucoup plus gênant : le Cloud Act américain

( **Blyx** un fournisseur soumis au droit US 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.

Ce n'est pas une théorie de forum. Le 10 juin 2025, devant la commission d'enquête du Sénat français, le directeur juridique de Microsoft France a été interrogé sous serment sur la possibilité de garantir que les données de citoyens français ne soient jamais transmises aux autorités américaines. Sa réponse : « **non, je ne peux pas le garantir** ». Et si l'entreprise y est contrainte, elle fournit les données.

À nuancer quand même : il faut un mandat validé par un juge américain sur des motifs sérieux. Ce n'est pas une aspiration permanente et automatique. C'est un **risque juridique structurel**, pas une surveillance continue.

Le deuxième point de fragilité : le cadre légal qui autorise aujourd'hui les transferts UE → USA (le Data Privacy Framework) est toujours valide, mais attaqué. Après une décision de la Cour suprême américaine du 29 juin 2026 fragilisant l'indépendance de la FTC, l'association noyb de Max Schrems a saisi la Commission dès le 30 juin et prépare un « Schrems III » devant la Cour de justice de l'UE. Une opinion est attendue fin 2026 / début 2027.

Autrement dit : miser 100 % de ta stack IA sur une API américaine, c'est accepter une dépendance dont les règles peuvent changer sans toi.

Et le paradoxe Kimi, à connaître absolument : l'API et le chat en ligne de Moonshot sont **pires** que les API américaines sur ce plan. Leur politique de confidentialité indique que le contenu utilisateur sert à 

entraîner et optimiser nos modèles », sans opt-out produit décrit, et ne
n'ont pas un pays de stockage. Précédent concret : l'Italie a bloqué  DeepSeek le 30 janvier 2025 pour stockage de données en Chine
contraire au RGPD.

👉 **La souveraineté vient des poids que tu télécharges et que tu fais tourner chez toi. Jamais de leur cloud.**

3. Arbre de décision : quel scénario pour ta boîte ?

Trois questions, dans cet ordre.

Question 1 — Quel est ton volume mensuel de tokens ? En dessous de ~800 millions de tokens par mois, l'auto-hébergement d'un gros modèle ne se rentabilise pas. Reste sur une API, mais choisis-la bien (scénario C).

Question 2 — Tes données sont-elles réglementées ? Santé, données bancaires, secret industriel, données de défense, données RH sensibles → la question du transfert hors UE se pose formellement, avec analyse d'impact à la clé.

Question 3 — As-tu une équipe infra capable d'opérer un cluster GPU ? Si la réponse est non, le scénario A est hors de portée, quel que soit ton budget.

Scénario A — Le modèle complet chez toi (grand groupe)

✓ **Pour qui :** plus de 1,5 milliard de tokens/mois, équipe infra dédiée, contraintes réglementaires fortes.





Matériel : Kimi K3 pèse environ 1,4 To de poids et vise un



Blyx

se déploie sur ~64 accélérateurs. C'est du multi-nœuds, pas un serveur.



- ✓ **Ordre de grandeur** : un serveur 8× H200 coûte entre 320 000 et 420 000 \$ à l'achat (typiquement ~370 000 \$). Il en faut plusieurs.
- ✓ **À savoir** : le modèle Moonshot réellement hébergeable aujourd'hui est **Kimi K2.6** (~1 000 milliards de paramètres). En FP8 il tient sur **8× H200** (~1 To de poids) ; en quantification INT4, sur **8× H100** (~630 Go). Stacks officiellement supportées : vLLM, SGLang, KTransformers.
- ✓ **Verdict** : réservé aux structures qui ont déjà un datacenter.

Scénario B — GPU souverain loué en France (ETI)

- ✓ **Pour qui** : 800 M à 1,5 Md tokens/mois, exigence de souveraineté, pas envie d'acheter du matériel.
- ✓ **Principe** : tu loues des GPU chez un hébergeur français, tu y déploies toi-même les poids open source. Les données ne sortent jamais de ton périmètre contractuel.
- ✓ **Ordre de grandeur** : un déploiement européen d'environ 2 nœuds revient à **18 000–28 000 € par mois** tout compris (GPU, énergie, réseau, astreinte, supervision). Chez Scaleway, un H100 SXM est à 2,73 €/h.
- ✓ **Point de vigilance** : au 19 juillet 2026, **aucun hébergeur français ne propose Kimi en endpoint managé**. Ni OVHcloud, ni Scaleway. Tu loues du GPU nu et tu déploies toi-même.
- ✓ **Verdict** : le meilleur compromis souveraineté/coût pour une ETI



Scénario C — Modèle plus léger ou API souveraine (PME)



Blyx



Pour qui : la très grande majorité des PME.



- ✓ **Option 1 — Mistral Large 3** (sorti le 2 décembre 2025) : 675 milliards de paramètres (41 actifs), contexte 256K, **licence Apache 2.0** — c'est-à-dire zéro contrainte, même à grande échelle. Surtout : il tient sur **un seul nœud 8× H100**, soit environ deux fois moins de matériel que Kimi K2.6. Éditeur français, sous juridiction européenne, déployable en interne ou en air-gap.
- ✓ **Option 2 — API hébergée en France** : OVHcloud AI Endpoints (datacenter de Gravelines, zéro rétention, zéro entraînement) propose plus de 40 modèles ouverts. Scaleway propose ses Generative APIs depuis ses datacenters parisiens. Coût marginal, aucune infra à gérer.
- ✓ **Verdict** : dans 9 cas sur 10, c'est ici que se trouve la bonne réponse pour une PME. La souveraineté ne coûte pas 370 000 \$.

4. Comparatif des clouds souverains français

Acteur	Localisation	Ce qu'ils proposent	Kimi dispo ?	Certificatio
OVHcloud	Gravelines (FR)	AI Endpoints : 40+ modèles ouverts (Mistral, Llama 3.3 70B, Qwen, gpt-oss). GPU H100/H200 + bare metal	Non — demande client ouverte et priorisée	Demande SecNumClc déposée
Scaleway	Paris (FR)	Generative APIs + Managed/Dedicated	Non	—



Acteur	Localisation	Ce qu'ils proposent	Kimi dispo ?	Certification
		Inference (Mistral Large 3, Qwen, Llama). H100 SXM 2,73 €/h		
Outscale	France	IaaS / GPU pour héberger tes propres poids	Non (à déployer soi-même)	SecNumCloud 3.2
S3NS (Google + Thales)	France	Cloud de confiance « Premi3ns »	—	SecNumCloud 3.2 (déc. 2025)
Bleu (Microsoft + Orange + Capgemini)	France	Cloud de confiance	—	Disponibilité annoncée S 2026

SecNumCloud, c'est quoi ? Le référentiel de l'ANSSI (version 3.2, plus de 360 critères) qui atteste à la fois la sécurité technique **et** l'immunité aux lois extraterritoriales comme le Cloud Act. C'est le vrai standard si tu traites des données sensibles.

À noter : des endpoints Kimi hébergés en Europe existent (Nscale, EUrouter, Nebius), mais aucun n'est un acteur français qualifié SecNumCloud.

5. Checklist conformité — RGPD et AI Act

Échéance immédiate : le 2 août 2026. C'est la date d'application majoritairement des obligations de l'EU AI Act, dont la gouvernance des



- ✓ ☐ **Cartographier.** Quels outils IA tes équipes utilisent-elles réellement ? (Attention : la réponse honnête est toujours plus longue que la réponse officielle.)
- ✓ ☐ **Identifier les données envoyées.** Données personnelles ? Données clients ? Secret des affaires ?
- ✓ ☐ **Vérifier le mécanisme de transfert.** Pour chaque fournisseur hors UE : décision d'adéquation, certification Data Privacy Framework, ou clauses contractuelles types + analyse d'impact des transferts.
- ✓ ☐ **Signer un accord de sous-traitance (article 28 RGPD)** avec chaque fournisseur. Tu restes responsable de leurs défaillances.
- ✓ ☐ **Activer le Zero Data Retention** si ton fournisseur le propose (Anthropic et OpenAI le font sur demande, pour les clients approuvés).
- ✓ ☐ **Réaliser une AIPD** (analyse d'impact) dès qu'il s'agit de données RH, de santé ou de profilage.
- ✓ ☐ **Écrire une politique IA interne** et la diffuser. Voir section 6 : c'est le point le plus rentable de toute cette liste.
- ✓ ☐ **Documenter tes flux de données.** L'AI Act exige de pouvoir prouver où vont les données — un audit ne se satisfait pas d'une bonne intention.
- ✓ ☐ **Revoir la décision tous les 6 mois.** L'écosystème bouge vite : Moonshot a sorti K2.5, K2.6, K2.7 et K3 en six mois.



6. Le risque que tout le monde sous-estime : le



Avant d'investir 370 000 \$ dans un rack GPU, regarde ce chiffre.

Selon le rapport **IBM Cost of a Data Breach 2025** :

- ✓ Coût moyen d'une violation de données : **4,44 millions de dollars**
- ✓ Le **shadow AI** — l'usage non encadré d'outils IA par les salariés — est impliqué dans **20 % des violations**
- ✓ Surcoût moyen quand il est impliqué : **jusqu'à 670 000 \$**
- ✓ **63 %** des organisations touchées n'avaient pas de politique de gouvernance IA

Le précédent le plus parlant reste Samsung : en mai 2023, l'entreprise a interdit ChatGPT en interne après **trois fuites en une vingtaine de jours**, dont du code source lié aux semi-conducteurs et des notes de réunion.

Et l'écart reste béant : d'après l'enquête Salesforce 2026, **67 % des salariés utilisent l'IA au travail, mais seulement 18 % des organisations ont une politique de sécurité IA formelle.**

👉 Une politique interne écrite, plus un accès officiel à un outil correct, coûte quelques heures. C'est le meilleur retour sur investissement de ce guide.

7. Ils l'ont déjà fait

Ce ne sont pas des projets pilotes. Ce sont des déploiements en production.





L'État français — Albert. L'IA générative de l'État, développée par



Blyx

la NLLM, tourne sur des modèles ouverts (Llama 3, Mistral)



hébergés sur une infrastructure souveraine française. Plus de 70

projets publics, plus de **100 000 requêtes par semaine.**



Ministère des Armées x Mistral AI. Accord-cadre notifié le 16 décembre 2025, piloté par l'AMIAD. Modèles déployés sur matériel souverain, avec protection explicite contre les juridictions étrangères.



BNP Paribas. La banque a construit une plateforme interne « LLM as a Service » dans ses propres datacenters, avec Mistral et des modèles ouverts — explicitement pour éviter le shadow AI et les fuites de données confidentielles.



Health Data Hub. Une procédure a été lancée le 7 février 2026 pour migrer la plateforme des données de santé hors Microsoft Azure, vers un hébergeur qualifié SecNumCloud.



Apertus (Suisse). Un LLM entièrement ouvert développé par l'EPFL et l'ETH Zurich, déployé par Swisscom sur une plateforme souveraine.

Contrepoint honnête, parce qu'un guide qui ne montre qu'un côté ne sert à rien : d'après Menlo Ventures, l'open source ne représentait que **11 % du marché des LLM en entreprise en 2025, en recul par rapport à 19 % en 2024.** Le closed-source domine largement. La bascule vers l'open source est un choix stratégique assumé pour des raisons de souveraineté — pas encore une lame de fond.

8. La licence : ce que tu as le droit de faire



Les modèles Kimi de la génération K2 sont publiés sous **licence «**

Mistral MIT ». Concrètement, tu as tous les droits MIT : usage commercial, modification, fine-tuning sur tes données, redistribution, revente.



Une seule clause additionnelle : si ton produit dépasse **100 millions d'utilisateurs actifs mensuels** ou **20 millions de dollars de revenus mensuels**, tu dois afficher « Kimi K2 » dans l'interface de ton produit.

Pour une PME ou une ETI française, ces seuils ne sont jamais atteints.
Usage libre, zéro redevance, zéro contrainte d'affichage.

*Attention : la licence exacte de **Kimi K3** n'était pas encore publiée au 19 juillet 2026. Elle sort avec les poids, le 27 juillet. Tout indique une continuité avec la lignée K2, mais vérifie le texte avant de bâtir un produit dessus.*

Pour comparaison, **Mistral Large 3 est sous Apache 2.0** : aucune clause d'affichage, quel que soit ton volume. C'est plus permissif encore.

Récapitulatif — les 5 décisions à prendre

1. **Écris ta politique IA interne cette semaine.** C'est gratuit et ça traite 20 % des violations de données.
2. **Calcule ton volume réel de tokens par mois.** Sous 800 millions, l'auto-hébergement d'un gros modèle n'est pas rentable.



3. Choisis ton scénario : A (rack interne), B (GPU souverain loué), ou C (Moonshot léger / API française). Pour la plupart des PME, c'est C. 

4. Vérifie tes transferts de données avant le 2 août 2026.

5. Attends le 27 juillet pour juger Kimi K3 sur pièces : licence définitive et évaluations indépendantes.

Guide rédigé le 19 juillet 2026 à partir d'une recherche sur sources primaires (documentation Moonshot AI, Artificial Analysis, IBM, ANSSI, comptes rendus du Sénat). Les chiffres de coûts sont des ordres de grandeur de marché, pas des devis. L'écosystème évolue vite : vérifie les données avant toute décision d'achat.

Une question sur ton cas précis ? Réponds-moi directement en DM.

Blyx — experts IA & développement · blyx-agency.com

