AYI NEDJIMI Consultants

LLM On-Premise vs Cloud : Souveraineté et Performance

Catégorie : Intelligence Artificielle Lecture : 26 min Publié le : 13/02/2026 Auteur : Ayi NEDJIMI

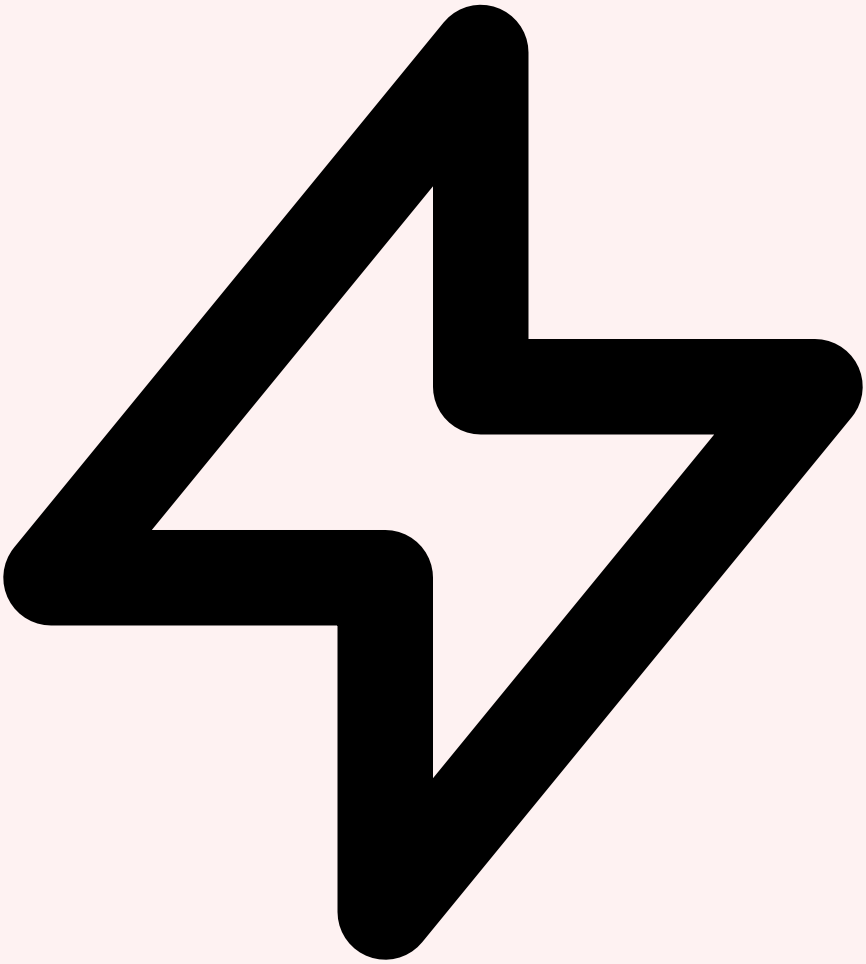
Guide complet comparant LLM on-premise vs cloud : souveraineté des données, performance GPU, coûts TCO, conformité RGPD/AI Act, architectures.

LLM On-Premise vs Cloud : Souveraineté et Performance constitue un enjeu majeur pour les professionnels de la sécurité informatique et les équipes techniques. Guide complet comparant LLM on-premise vs cloud : souveraineté des données, performance GPU, coûts TCO, conformité RGPD/AI Act, architectures. Ce guide détaillé sur ia llm on premise vs 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

1 Les Enjeux du Déploiement de LLM en 2026

L'année 2026 marque un tournant décisif dans la maturité des **Large Language Models** au sein des entreprises. Après une phase d'expérimentation massive entre 2023 et 2025, les organisations se trouvent confrontées à une question structurante : **où et comment déployer leurs modèles de langage en production** ? Ce choix, loin d'être purement technique, engage des dimensions stratégiques — souveraineté des données, conformité réglementaire, maîtrise des coûts et performance opérationnelle — qui détermineront la capacité des entreprises à exploiter l'IA générative de manière durable et responsable. Le marché mondial de l'inférence LLM a dépassé les **85 milliards de dollars** en 2025, et les projections pour 2026 indiquent une croissance de 40 % tirée par la généralisation des cas d'usage en production : assistance client, génération documentaire, analyse juridique, code assisté et aide à la décision stratégique.



Le spectre des options de déploiement

Le paysage du déploiement LLM s'est considérablement structuré et se décline désormais en trois grandes familles. Le **cloud public**, porté par OpenAI, Anthropic, Google et AWS Bedrock, offre l'accès aux modèles les plus puissants via API avec un time-to-market quasi instantané, mais implique l'envoi de données sensibles vers des infrastructures tierces. Le **déploiement on-premise**, rendu viable par la démocratisation des modèles open-weight (Llama 3.1, Mistral Large, Qwen 2.5, DeepSeek-V3) et des frameworks d'inférence optimisés (vLLM, TGI, TensorRT-LLM), garantit un contrôle total sur les données mais exige des investissements matériels significatifs et une expertise technique pointue. Entre ces deux extrêmes, les **architectures hybrides** combinent cloud et on-premise selon la sensibilité des données et les exigences de latence, offrant un compromis pragmatique que de plus en plus d'entreprises adoptent en 2026.

Notre avis d'expert

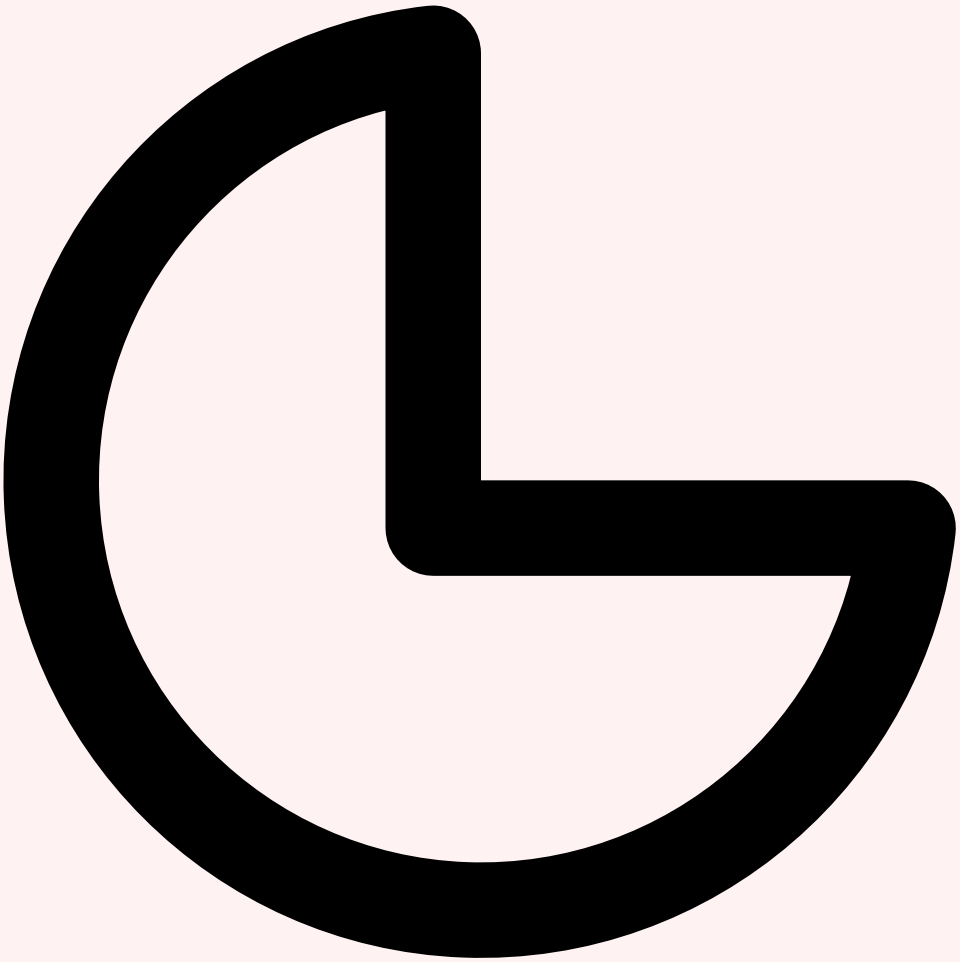
L'IA responsable n'est pas un luxe — c'est une nécessité opérationnelle. Nos audits révèlent que 70% des déploiements IA en entreprise manquent de mécanismes de détection des biais et de garde-fous contre les injections de prompt. Il est temps d'intégrer la sécurité dès la conception des pipelines ML.

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

Le contexte réglementaire européen

Le cadre juridique européen ajoute une couche de complexité déterminante au choix de déploiement. L'**AI Act**, entré en application progressive depuis août 2024, impose des exigences spécifiques selon le niveau de risque des systèmes d'IA : transparence, documentation technique, gestion des biais, traçabilité des décisions et supervision humaine. Pour les systèmes classés à **haut risque** — notamment dans les domaines RH, santé, justice et services financiers —, le règlement exige une documentation exhaustive des données d'entraînement, des tests de robustesse et un monitoring continu en

production. Le **RGPD**, toujours en vigueur, contraint la localisation et le traitement des données personnelles, rendant problématique l'utilisation d'APIs cloud hébergées hors de l'Union européenne, même avec les clauses contractuelles types post-Schrems II. La **directive NIS2**, applicable depuis octobre 2024, renforce les obligations de cybersécurité pour les opérateurs de services essentiels et importants, ce qui inclut désormais les infrastructures IA critiques.



Les critères de décision stratégiques

La décision entre cloud, on-premise et hybride repose sur une matrice multicritères que chaque organisation doit évaluer selon son contexte propre. Les **critères techniques** incluent la latence cible (temps de réponse de premier token, TTFT), le débit nécessaire (tokens par seconde), la taille des modèles visés et les fenêtres de contexte requises. Les **critères de gouvernance** englobent la classification des données traitées, les obligations réglementaires sectorielles, les politiques de rétention et le droit à l'oubli. Les **critères économiques** portent sur le TCO (Total Cost of Ownership) à 3 et 5 ans, les CAPEX vs OPEX, la prévisibilité budgétaire et l'élasticité de la demande. Enfin, les **critères organisationnels**

évaluent la maturité de l'équipe MLOps, la capacité à recruter des experts GPU et la volonté de l'entreprise d'internaliser cette compétence stratégique. L'erreur la plus fréquente en 2026 reste de réduire ce choix à une simple comparaison de coûts unitaires d'inférence, en ignorant les coûts cachés d'intégration, de maintenance et de conformité qui représentent souvent 40 à 60 % du TCO réel.

- ▹**Performance vs contrôle** — Les APIs cloud offrent les modèles les plus performants (GPT-4o, Claude Opus, Gemini Ultra) mais sans maîtrise sur l'infrastructure sous-jacente ni garantie de reproductibilité
- ▹**Souveraineté vs agilité** — Le on-premise garantit la localisation des données mais allonge le time-to-market de 3 à 6 mois par rapport à une intégration cloud
- ▹**CAPEX vs OPEX** — Un cluster GPU on-premise représente un investissement initial de 500K à 2M EUR mais devient rentable au-delà d'un certain seuil de requêtes mensuelles
- ▹**Scalabilité vs prévisibilité** — Le cloud s'adapte instantanément aux pics de charge, mais les factures peuvent devenir imprévisibles sans governance FinOps rigoureuse

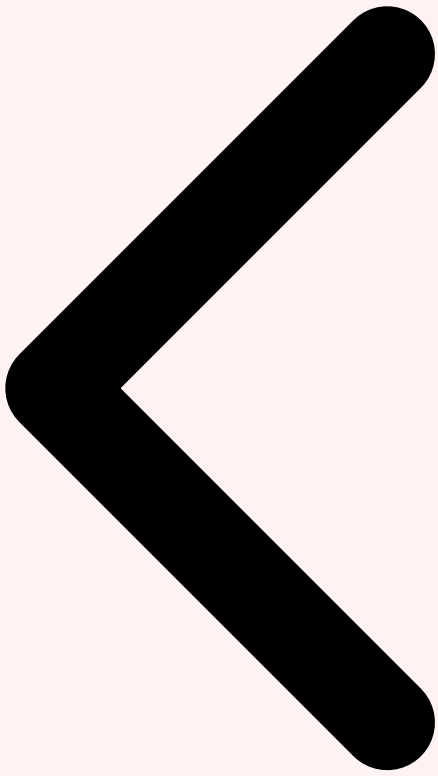
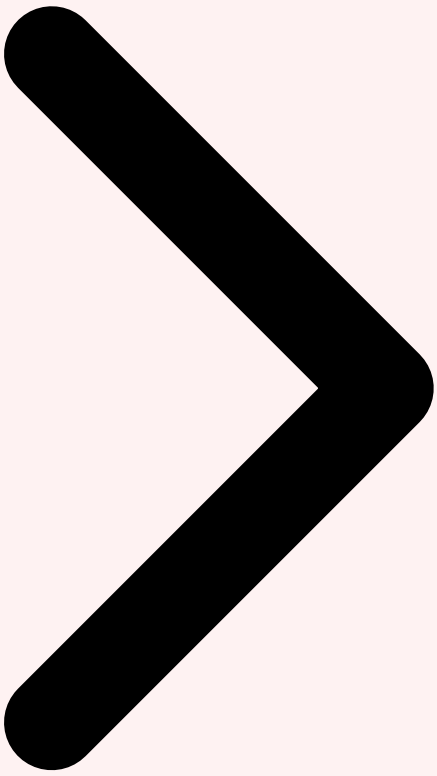
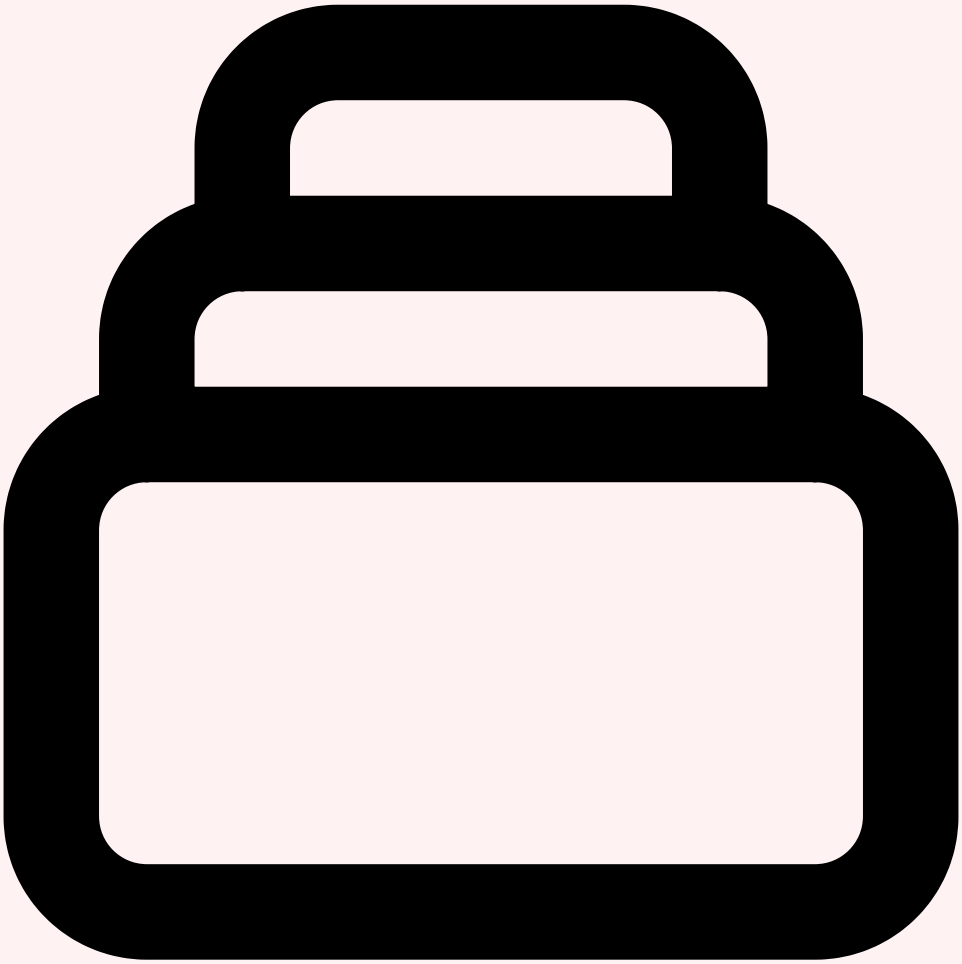


Table des Matières Enjeux Déploiement Cloud APIs



2 Déploiement Cloud : APIs et Managed Services

Le déploiement cloud des LLM constitue la voie la plus directe pour intégrer l'intelligence artificielle générative dans les processus métier. En 2026, l'écosystème s'est consolidé autour de deux modèles complémentaires : les **APIs de modèles propriétaires** (OpenAI, Anthropic, Google, Cohere) et les **plateformes d'inférence managées** (AWS Bedrock, Azure OpenAI Service, Google Vertex AI, OVHcloud AI Endpoints) qui proposent à la fois des modèles propriétaires et open-weight. L'avantage fondamental du cloud réside dans l'**élimination complète de la complexité infrastructure** : pas de GPU à approvisionner, pas de drivers CUDA à maintenir, pas de clusters Kubernetes à orchestrer. L'équipe d'ingénierie se concentre exclusivement sur la logique applicative — prompt engineering, chaînes RAG, pipelines d'évaluation — tandis que le fournisseur cloud gère l'optimisation de l'inférence, le scaling horizontal et la haute disponibilité.



APIs propriétaires : puissance et simplicité

Les APIs propriétaires offrent l'accès aux **modèles les plus performants** du marché. OpenAI avec GPT-4o et la série o3 de modèles de raisonnement, Anthropic avec Claude Opus 4 et son contexte de 200K tokens, Google avec Gemini 2.0 Ultra et ses capacités multimodales natives — ces modèles frontier surpassent systématiquement les alternatives open-weight sur les benchmarks complexes de raisonnement, de codage et d'analyse multilingue. L'intégration technique se réduit souvent à quelques lignes de code via des SDKs officiels robustes, avec des fonctionnalités avancées prêtes à l'emploi : **function calling**, **structured outputs** (JSON mode), **vision** (analyse d'images et documents), **streaming** et **batch processing**. En 2026, les providers cloud proposent également des garanties de SLA de 99,9 % voire 99,95 %, avec des latences TTFT (Time To First Token) moyennes de 100 à 250 ms selon le modèle et la région. Pour approfondir, consultez Reinforcement Learning Appliqué à la Cybersécurité.

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

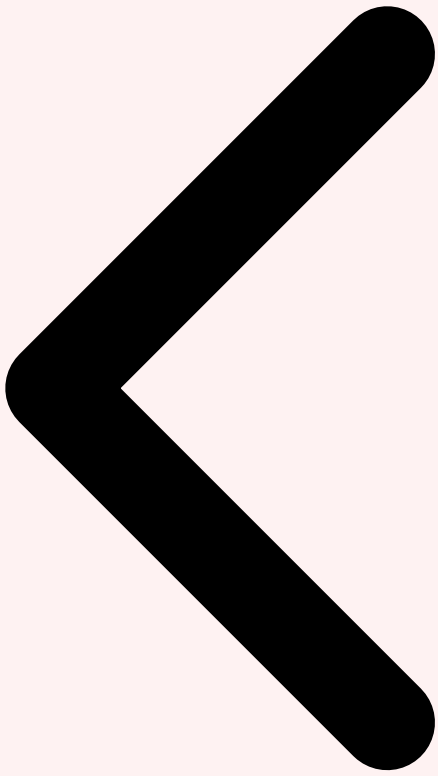
Provider	Service	Modèles Phares	Tarif Input/Output (1M tokens)	Data Residency EU
OpenAI	API directe	GPT-4o, o3-mini	$2.50 / $10.00	Partiel (EU via Azure)
Anthropic	API directe	Claude Opus 4, Sonnet	$15.00 / $75.00	Via AWS/GCP EU
AWS	Bedrock	Claude, Llama, Mistral	Variable par modèle	eu-west-1/3, eu-central-1
Azure	OpenAI Service	GPT-4o, Phi-3	Aligné sur OpenAI	France Central, West EU
Google	Vertex AI	Gemini 2.0, PaLM	$1.25 / $5.00 (Flash)	europe-west1/4/9
OVHcloud	AI Endpoints	Mistral, Llama	Compétitif	Gravelines, Strasbourg

Risques et limites du tout-cloud

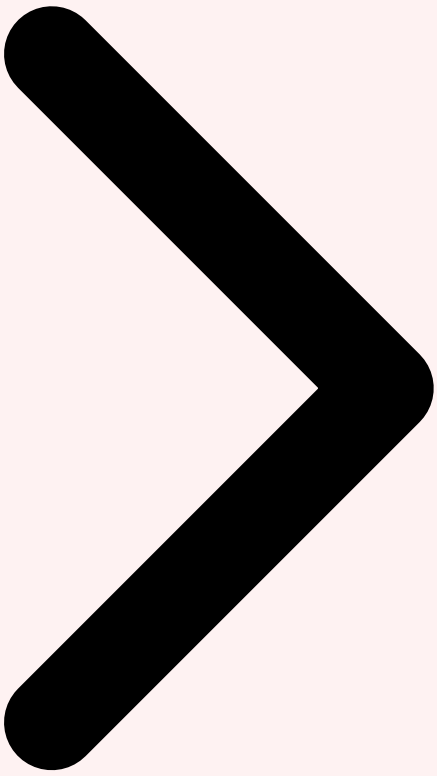
Malgré ses avantages indéniables, le déploiement cloud présente des risques structurels que les organisations doivent évaluer avec rigueur. Le premier est la **dépendance fournisseur** (vendor lock-in) : les prompts optimisés pour GPT-4o ne fonctionnent pas de manière identique avec Claude ou Gemini, les intégrations de function calling diffèrent entre providers, et les pipelines RAG doivent être réadaptés lors d'un changement de fournisseur. Le deuxième risque est la **perte de contrôle sur les données** : même avec les engagements contractuels de non-utilisation des données pour l'entraînement (opt-out), les requêtes transitent par des infrastructures tierces, sont journalisées, et potentiellement soumises à des juridictions extra-européennes via le Cloud Act américain ou des dispositifs équivalents. Le troisième risque concerne la **prévisibilité budgétaire** : le modèle pay-per-

token, attractif pour les phases de prototypage, peut générer des factures exponentielles en production lorsque le volume de requêtes augmente de manière organique, avec des coûts qui peuvent quadrupler en quelques mois sans gouvernance FinOps stricte.

- ▹**Latence réseau incompressible** — Chaque requête API ajoute 30 à 150 ms de latence réseau, problématique pour les use cases temps réel (chatbot, code completion, trading)
- ▹**Rate limiting et quotas** — Les providers imposent des limites de débit (tokens/min, requêtes/min) qui peuvent devenir bloquantes en phase de scaling production
- ▹**Versions de modèles non maîtrisées** — Les providers mettent à jour ou déprécient leurs modèles sans préavis suffisant, causant des régressions dans les pipelines de production
- ▹**Disponibilité et incidents** — Les pannes de services cloud (OpenAI a connu 47 incidents majeurs en 2025) impactent directement toute la chaîne applicative sans possibilité de failover local

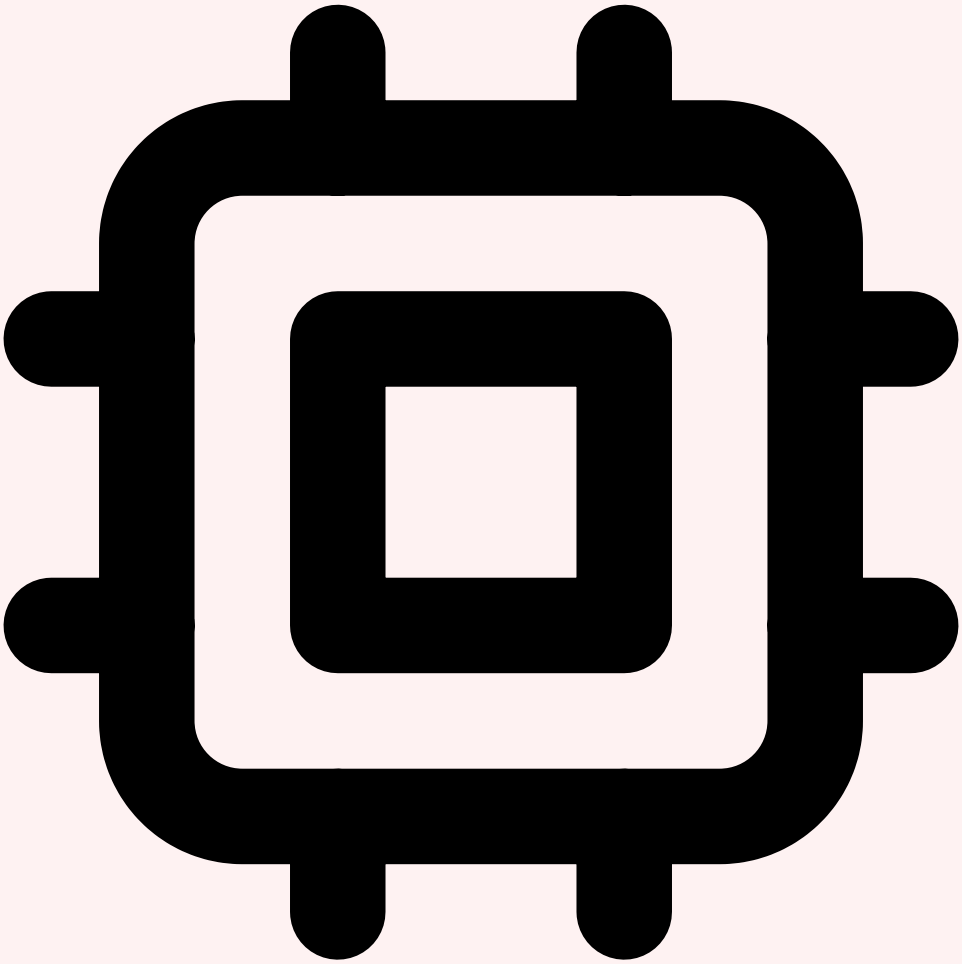


Enjeux Déploiement Cloud APIs On-Premise



3 Déploiement On-Premise : Contrôle Total

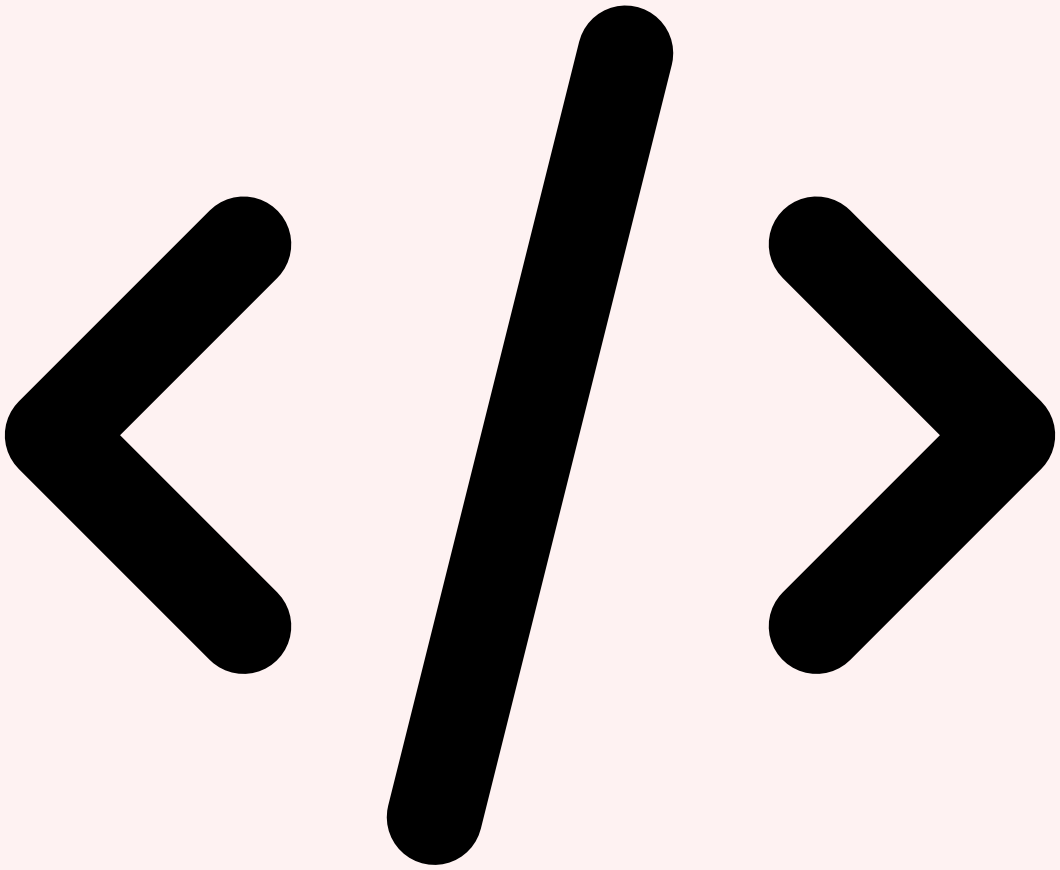
Le déploiement on-premise des LLM a connu une transformation radicale entre 2023 et 2026. Ce qui relevait encore de l'exploit technique il y a trois ans — faire tourner un modèle de 70 milliards de paramètres sur une infrastructure locale — est devenu un processus industrialisé grâce à la convergence de trois facteurs : la **démocratisation des modèles open-weight** de qualité frontier, la **maturation des frameworks d'inférence** optimisés et la **disponibilité croissante du hardware GPU** de nouvelle génération. En 2026, un cluster de 8 GPU NVIDIA H200 (141 Go HBM3e chacun) peut servir en inférence un modèle de 405 milliards de paramètres quantifié en FP8 avec des performances comparables aux APIs cloud, tout en maintenant les données strictement dans le périmètre de l'organisation. Cette capacité transforme fondamentalement l'équation souveraineté-performance qui rendait le on-premise prohibitif auparavant.



Le hardware GPU : état de l'art 2026

Le choix du matériel GPU constitue la décision la plus structurante du déploiement on-premise. NVIDIA domine toujours le marché de l'inférence LLM en 2026, mais le paysage s'est diversifié. La gamme **H200** (141 Go HBM3e, 4,8 TB/s de bande passante mémoire) représente le choix de référence pour l'inférence de modèles de grande taille, offrant un gain de 50 à 90 % de throughput par rapport au H100 sur les workloads LLM grâce à sa mémoire HBM3e étendue. La **B200** (192 Go HBM3e, architecture Blackwell), disponible depuis fin 2025, pousse les performances encore plus loin avec un moteur de second génération pour le FP4 et le FP8, permettant de servir des modèles de 405B paramètres sur seulement 4 GPU en quantification FP4. Pour les budgets plus contraints, les **GPU AMD MI300X** (192 Go HBM3, 5,3 TB/s) représentent une alternative crédible avec un rapport performance-prix supérieur de 20 à 30 % à NVIDIA sur certains workloads d'inférence, bien que l'écosystème logiciel ROCm reste moins mature que CUDA.

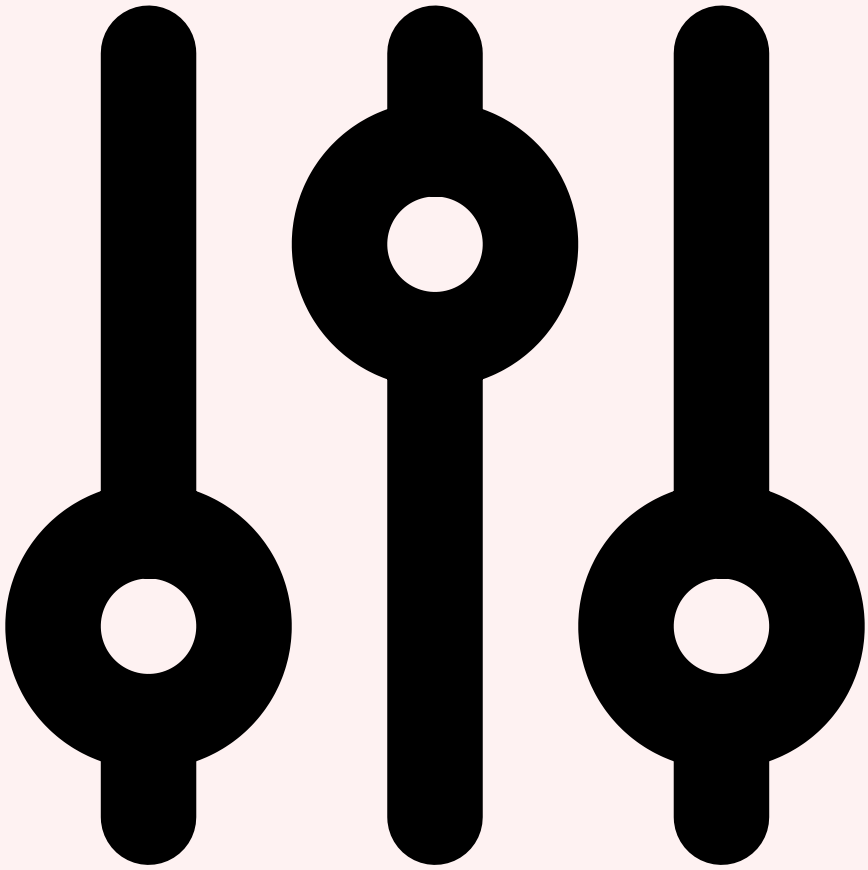
GPU	VRAM	Bandwidth	Modèle Max (FP8)	Prix unitaire	Cas d'usage
NVIDIA H100 SXM	80 Go HBM3	3,35 TB/s	70B (×1) / 405B (×8)	~25 000 EUR	Production standard
NVIDIA H200 SXM	141 Go HBM3e	4,8 TB/s	120B (×1) / 405B (×4)	~35 000 EUR	Grands modèles
NVIDIA B200	192 Go HBM3e	8 TB/s	405B (×4, FP4)	~45 000 EUR	Frontier on-prem
AMD MI300X	192 Go HBM3	5,3 TB/s	120B (×1) / 405B (×4)	~22 000 EUR	Alternative coût
Intel Gaudi 3	128 Go HBM2e	3,7 TB/s	70B (×1)	~15 000 EUR	Workloads spécifiques



Stack logiciel d'inférence on-premise

La stack logicielle d'inférence a atteint un niveau de maturité industriel en 2026. **vLLM** s'est imposé comme le standard de facto pour l'inférence open-source, grâce à son implémentation du **PagedAttention** qui optimise la gestion de la mémoire GPU en traitant l'attention comme un système de pagination virtuelle, permettant d'augmenter le throughput de 2 à 4x par rapport à une implémentation naïve. vLLM supporte nativement le **continuous batching**, le speculative decoding, le prefix caching et le tensor parallelism multi-GPU, avec une API compatible OpenAI qui facilite la migration depuis les APIs cloud. **TensorRT-LLM** de NVIDIA offre des performances supérieures de 15 à 30 % à vLLM sur les GPU NVIDIA grâce à des optimisations kernel spécifiques, mais au prix d'une moindre flexibilité et d'une dépendance à l'écosystème NVIDIA. **SGLang**, développé par l'équipe de Berkeley, se positionne comme le challenger avec des innovations sur le structured decoding et le RadixAttention pour le prefix caching.

```
# Déploiement vLLM avec Docker et GPU NVIDIA
docker run --gpus all -d \
  --name vllm-server \
  -p 8000:8000 \
  -v /models:/models \
  vllm/vllm-openai:latest \
  --model /models/Meta-Llama-3.1-70B-Instruct \
  --tensor-parallel-size 2 \
  --max-model-len 32768 \
  --gpu-memory-utilization 0.92 \
  --enable-prefix-caching \
  --quantization fp8

# Configuration Kubernetes avec GPU Operator
apiVersion: apps/v1
kind: Deployment
metadata:
  name: vllm-llama-70b
spec:
  replicas: 2
  template:
    spec:
      containers:
      - name: vllm
        resources:
          limits:
            nvidia.com/gpu: 2
        env:
        - name: VLLM_ATTENTION_BACKEND
          value: FLASH_ATTN
```

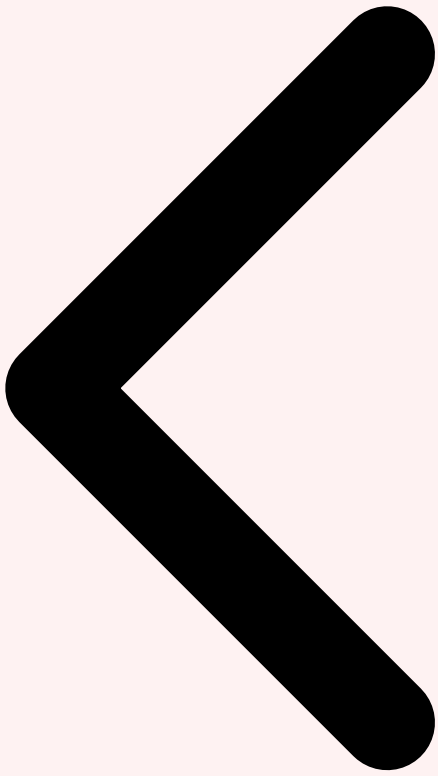


Les modèles open-weight de grade production

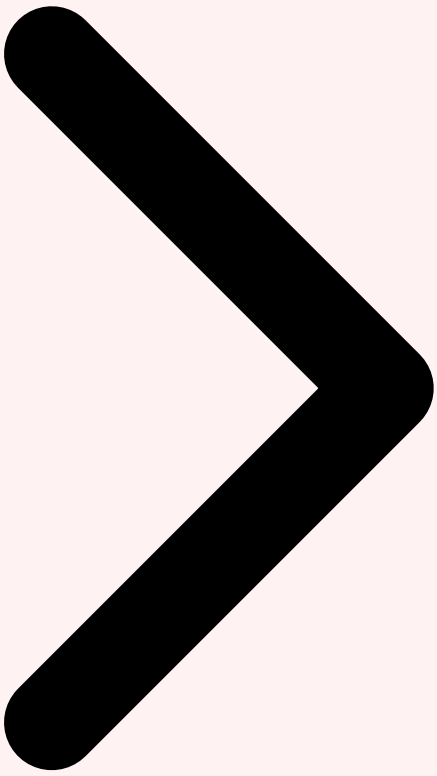
La qualité des modèles open-weight a atteint un niveau qui les rend pleinement viables pour la production on-premise. **Meta Llama 3.1 405B** offre des performances comparables à GPT-4 sur la majorité des benchmarks, avec une licence commerciale permissive. **Mistral Large 2** (123B paramètres) excelle sur les tâches multilingues européennes et propose un support commercial via Mistral AI. **DeepSeek-V3** (671B paramètres, architecture MoE avec 37B actifs) représente une percée en efficacité computationnelle, rivalisant avec les modèles frontier tout en nécessitant une fraction de la puissance GPU grâce à son architecture Mixture of Experts. **Qwen 2.5 72B** d'Alibaba se distingue sur les tâches de coding et de mathématiques. Ces modèles, combinés à des techniques de quantification avancées (GPTQ, AWQ, GGUF pour les formats les plus courants, et FP8 natif sur les GPU H100/H200), permettent de servir des modèles de 70B paramètres sur un seul GPU H200 avec une qualité quasi identique au modèle original en FP16, rendant le on-premise accessible à un spectre beaucoup plus large d'organisations.

- ▹**Reproducibilité garantie** — Le modèle et ses poids sont versionnés localement, éliminant le risque de changement unilatéral par un provider cloud et assurant la stabilité des pipelines

- ▹**Fine-tuning souverain** — La possibilité d'adapter le modèle sur des données propriétaires sans les exposer à un tiers, via LoRA, QLoRA ou full fine-tuning sur l'infrastructure locale
- ▹**Latence ultra-faible** — L'inférence locale élimine la latence réseau, atteignant un TTFT de 5 à 15 ms pour les modèles optimisés, critique pour le code completion et les agents autonomes
- ▹**Coût marginal décroissant** — Une fois l'infrastructure amortie, le coût par token tend vers zéro, rendant le on-premise économiquement supérieur au-delà de 10 à 50M tokens/jour selon la configuration



Cloud APIs On-Premise Souveraineté



4 Souveraineté des Données et Conformité

La souveraineté des données est devenue le critère le plus déterminant dans le choix d'architecture de déploiement LLM pour les entreprises européennes en 2026. Au-delà du simple concept de **localisation géographique des données**, la souveraineté engage trois dimensions complémentaires : la **souveraineté technique** (maîtrise de l'infrastructure et du code), la **souveraineté juridique** (contrôle du cadre légal applicable aux données) et la **souveraineté opérationnelle** (capacité à opérer indépendamment de fournisseurs tiers). L'actualité réglementaire de 2025-2026 a considérablement renforcé ces exigences : l'entrée en vigueur progressive de l'AI Act, le renforcement des contrôles CNIL sur les traitements IA, et les premières sanctions européennes liées au transfert de données personnelles vers des modèles d'IA hébergés hors UE ont créé un impératif juridique concret. Pour les secteurs régulés — banque, assurance, santé, défense, énergie — la non-conformité peut entraîner des amendes allant jusqu'à 35 millions d'euros ou 7 % du chiffre d'affaires mondial au titre de l'AI Act, auxquelles s'ajoutent les 20 millions d'euros ou 4 % du CA mondial prévus par le RGPD.

Le cadre RGPD appliqué aux LLM

L'application du RGPD aux systèmes LLM soulève des questions juridiques spécifiques que les entreprises doivent anticiper dans leur choix d'architecture. La première concerne la **base légale du traitement** : lorsqu'un LLM traite des données personnelles — noms dans des contrats, données RH, informations clients —, le responsable de traitement doit justifier d'une base légale parmi celles prévues à l'article 6 du RGPD (consentement, intérêt légitime, exécution contractuelle). En cas d'utilisation d'une API cloud, la question du **transfert de données vers un pays tiers** se pose immédiatement : les clauses contractuelles types (SCCs) et les Data Processing Addendum (DPA) des providers cloud offrent un cadre juridique, mais leur solidité a été fragilisée par l'arrêt Schrems II et reste contestée par certaines autorités de protection des données européennes. Le **droit à l'effacement** (article 17) est particulièrement complexe à mettre en oeuvre avec les LLM : si un modèle a été fine-tuné sur des données personnelles, la suppression de ces données du dataset d'entraînement n'efface pas mécaniquement l'information du modèle, nécessitant des techniques de machine unlearning encore expérimentales. Pour approfondir, consultez RAG Architecture | Guide.

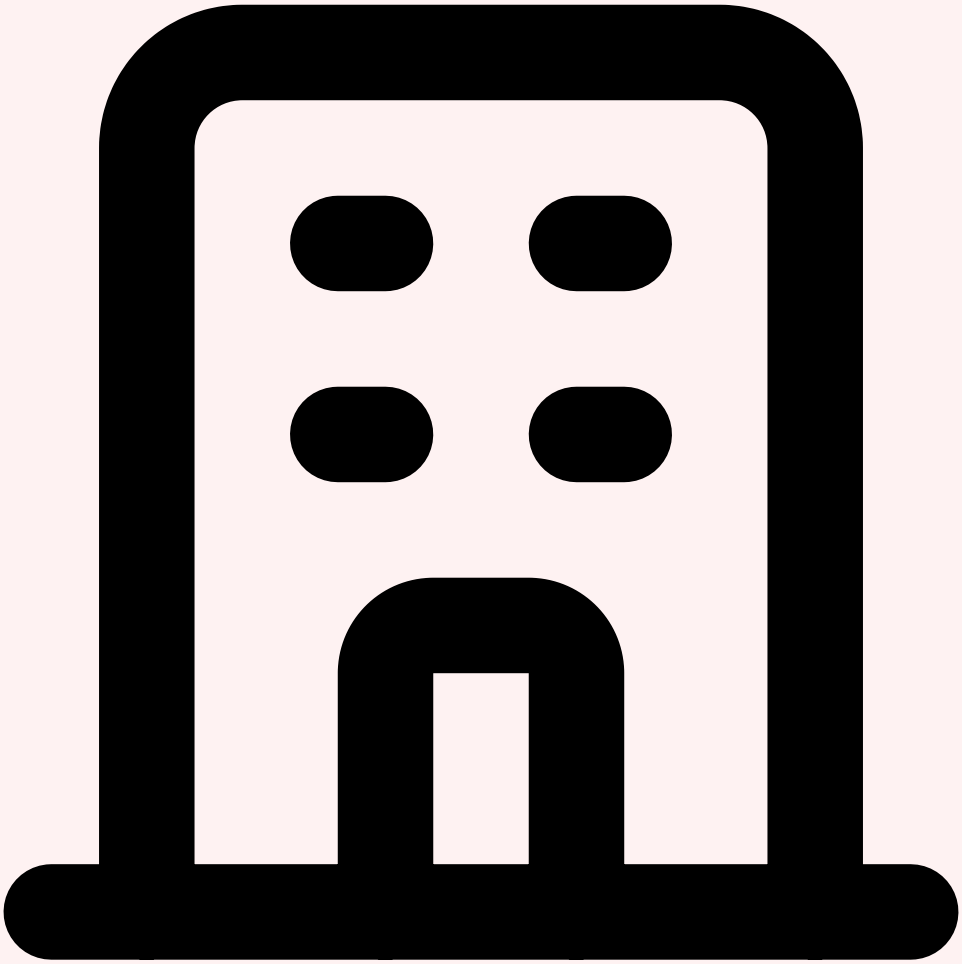
Point juridique clé : La CNIL a publié en 2025 des recommandations spécifiques aux systèmes d'IA générative, exigeant une **analyse d'impact relative à la protection des données (AIPD)** pour tout déploiement de LLM traitant des données personnelles à grande échelle. Cette AIPD doit documenter la finalité du traitement, les catégories de données, les mesures de minimisation, les mécanismes de filtrage des données personnelles dans les prompts, et les procédures de réponse aux demandes d'exercice des droits. En cas de déploiement cloud, l'AIPD doit également évaluer les risques liés au transfert transfrontalier et les mesures techniques supplémentaires mises en oeuvre (chiffrement de bout en bout, pseudonymisation avant envoi).

AI Act : obligations par niveau de risque

L'AI Act européen structure ses exigences selon une approche par les risques qui impacte directement le choix d'architecture de déploiement. Les systèmes d'IA à **risque inacceptable** (score social, manipulation subliminale) sont interdits, ce qui ne concerne pas les LLM d'usage général. Les systèmes à **haut risque** — qui incluent les LLM utilisés dans le recrutement, l'évaluation de crédit, la justice prédictive, les dispositifs médicaux et les

infrastructures critiques — doivent satisfaire des exigences de documentation technique, de qualité des données d'entraînement, de traçabilité des décisions, de robustesse et de supervision humaine. Pour ces systèmes, le déploiement on-premise offre un avantage structurel : il permet de **documenter intégralement la chaîne de traitement**, de maintenir un registre auditable des requêtes et réponses, et de garantir que les données d'évaluation et de test restent sous contrôle. Les systèmes d'IA à **usage général** (GPAI), catégorie qui inclut les LLM foundation models, sont soumis à des obligations de transparence et de documentation technique qui incombent principalement au fournisseur du modèle, mais le déployeur reste responsable de l'utilisation conforme du système dans son contexte spécifique.

Exigence	On-Premise	Cloud EU	Cloud US
Localisation données UE	Garanti	Garanti (région EU)	Non conforme
Protection Cloud Act	Non applicable	Risque résiduel	Exposé
Audit technique complet	Accès total	Limité (SLA)	Très limité
Traçabilité AI Act	Logs complets	Logs API partiels	Logs API partiels
Droit à l'effacement	Contrôle direct	Dépend du DPA	Complexe
Qualification SecNumCloud	Possible	3 providers FR qualifiés	Impossible

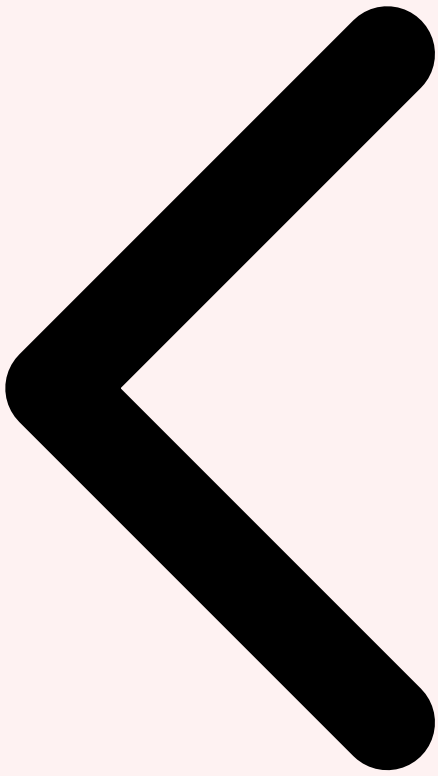


SecNumCloud et cloud souverain français

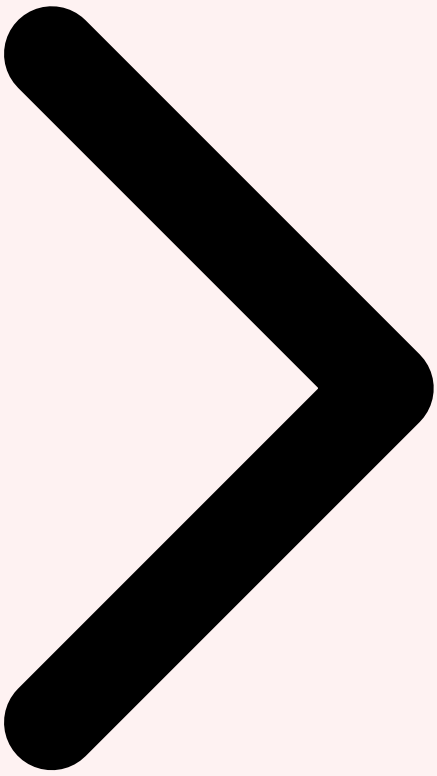
La qualification **SecNumCloud** de l'ANSSI constitue le référentiel de confiance le plus exigeant pour les services cloud en France. En 2026, trois fournisseurs cloud français ont obtenu cette qualification pour leurs services d'infrastructure et de plateforme : **OVHcloud**, **Outscale** (filiale de Dassault Systèmes) et **Scaleway**. Ces providers proposent désormais des services GPU managés permettant le déploiement de LLM dans un environnement qualifié SecNumCloud, combinant les avantages du cloud (élasticité, PaaS) avec les garanties de souveraineté exigées par l'État et les opérateurs d'importance vitale (OIV). La doctrine « cloud au centre » de l'État français, mise à jour en 2025, impose l'utilisation de services qualifiés SecNumCloud pour tous les traitements de données sensibles de l'administration, ce qui inclut désormais explicitement les systèmes d'IA générative. Pour les entreprises privées des secteurs régulés (banque, santé, énergie), cette qualification devient un prérequis contractuel exigé par les régulateurs sectoriels (ACPR, HAS, CRE) pour l'utilisation de LLM en production sur des données clients ou patients.

- ▹**Classification des données** — Établir une taxonomie claire (public, interne, confidentiel, secret) et mapper chaque catégorie sur l'architecture de déploiement autorisée

- ▹**DLP avant inférence** — Déployer des garde-fous de Data Loss Prevention qui filtrent les données personnelles et sensibles avant envoi vers une API cloud
- ▹**Chiffrement de bout en bout** — Utiliser le chiffrement TLS 1.3 pour le transit et le chiffrement AES-256 pour le stockage des logs de requêtes et réponses
- ▹**Audit trail immutable** — Journaliser chaque requête LLM dans un système de logs immuables (append-only) avec horodatage, identifiant utilisateur, hash du prompt et classification de sensibilité



On-Premise Souveraineté Comparatif Coûts



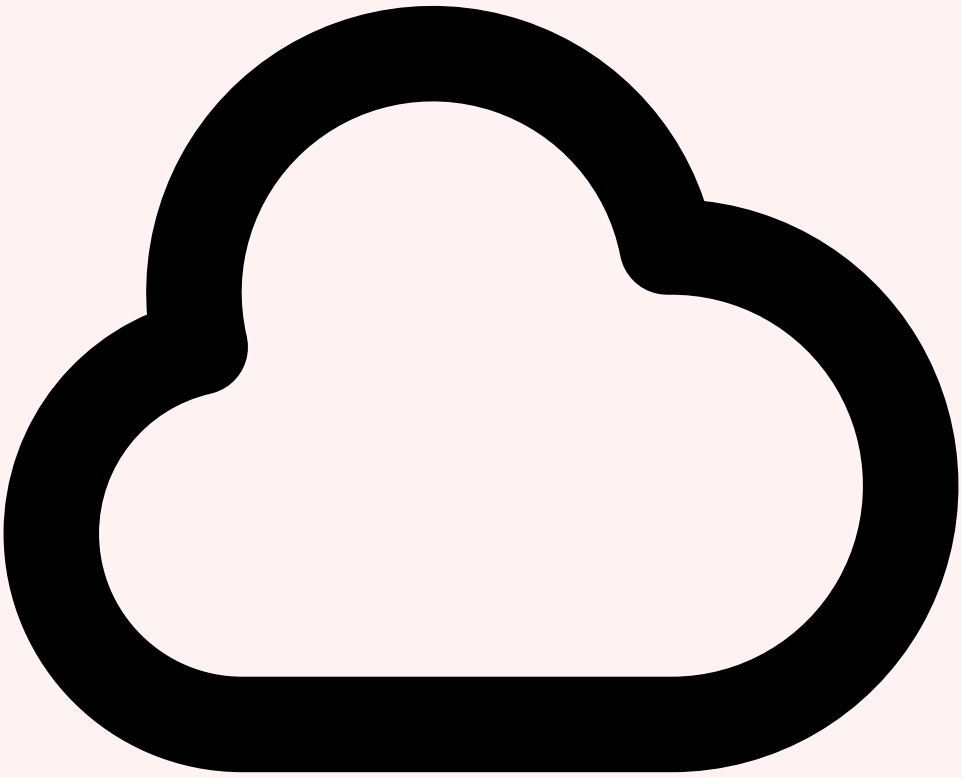
5 Comparatif des Coûts : TCO Cloud vs On-Premise

L'analyse du **Total Cost of Ownership** (TCO) constitue l'exercice le plus complexe et le plus critique du choix d'architecture LLM. Les comparaisons superficielles — coût par token cloud vs coût d'amortissement GPU — masquent la réalité d'un calcul qui doit intégrer des dizaines de variables, certaines évidentes (prix des GPU, tarifs API) et d'autres souvent négligées (coût de l'électricité, climatisation, personnel MLOps, coût d'opportunité du time-to-market). L'erreur la plus fréquente est de comparer le coût marginal d'un token cloud au coût moyen d'un token on-premise sans prendre en compte les **coûts fixes** (infrastructure, réseau, stockage, licences), les **coûts d'exploitation** (personnel, énergie, maintenance matérielle) et les **coûts cachés** (formation, recrutement, coût d'indisponibilité, dette technique). Voici une méthodologie structurée pour un calcul de TCO rigoureux sur un horizon de 3 ans, horizon standard pour l'amortissement d'un cluster GPU.

TCO On-Premise : décomposition détaillée

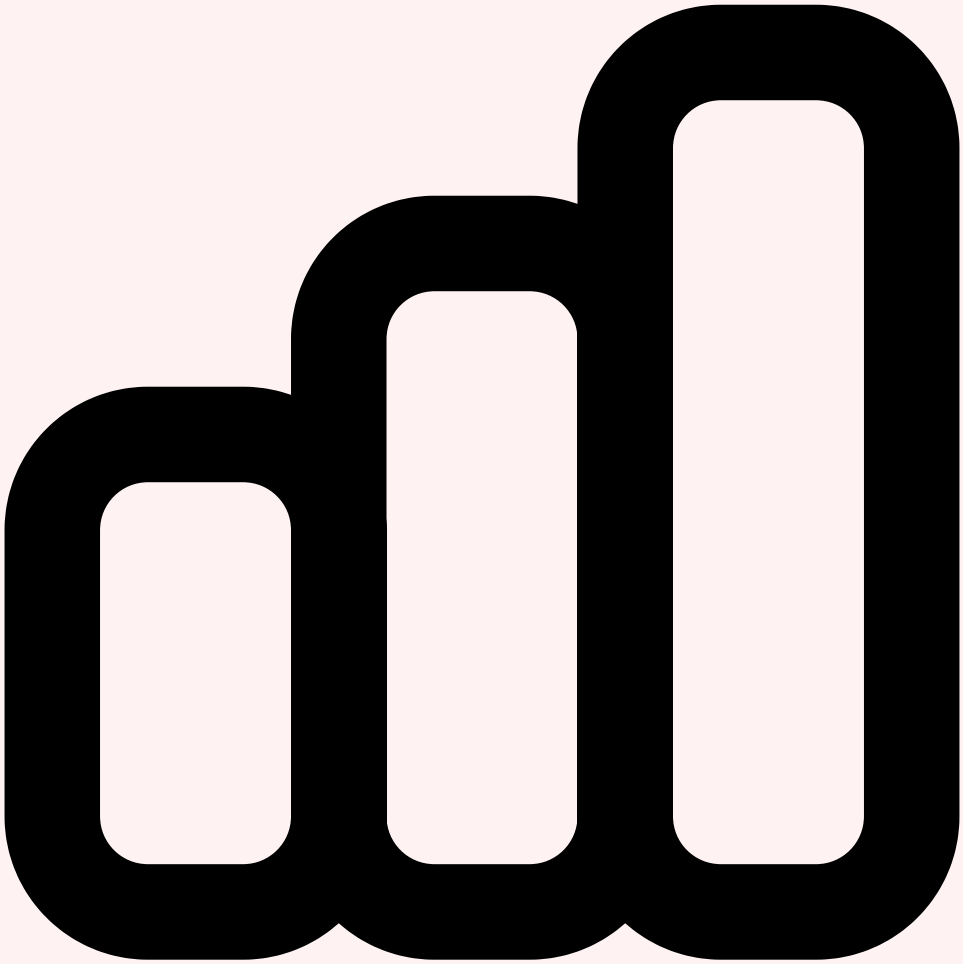
Le TCO on-premise se décompose en quatre catégories majeures. Les **coûts d'acquisition** (CAPEX) représentent l'investissement initial : pour un cluster de production standard de 8 GPU NVIDIA H200 avec serveur DGX H200, le coût se situe entre 350 000 et 450 000 EUR pour le matériel seul, auquel s'ajoutent 30 000 à 80 000 EUR pour l'infrastructure réseau (switches InfiniBand 400 Gb/s), 15 000 à 40 000 EUR pour le stockage NVMe rapide (modèles + cache KV), et 20 000 à 50 000 EUR pour l'installation physique (alimentation électrique renforcée, climatisation, baie rack). Les **coûts d'exploitation** (OPEX) annuels comprennent l'électricité (un DGX H200 consomme environ 10,2 kW sous charge, soit 20 000 à 30 000 EUR/an en France selon le tarif), la maintenance matérielle (contrat support NVIDIA à 15-20 % du prix d'achat par an), et les licences logicielles (NVIDIA AI Enterprise à 4 500 EUR/GPU/an, ou alternatives open-source gratuites). Le poste le plus significatif est souvent le **coût humain** : un ingénieur MLOps senior coûte entre 70 000 et 110 000 EUR/an en France (salaire chargé), et une équipe minimale viable pour opérer un cluster GPU en production 24/7 nécessite au moins 2 à 3 ETP (ingénieur MLOps, ingénieur infrastructure, DevOps/SRE à temps partiel), soit 180 000 à 300 000 EUR/an.

```
# Calculateur TCO On-Premise (3 ans) — Cluster 8x H200

CAPEX (investissement initial)
  Serveur DGX H200 (8x GPU)        380 000 EUR
  Infrastructure réseau InfiniBand    50 000 EUR
  Stockage NVMe 30 TB                 25 000 EUR
  Installation + aménagement           35 000 EUR
  Total CAPEX                         490 000 EUR

OPEX annuel (exploitation)
  Électricité (10.2 kW × 8760h)       25 000 EUR/an
  Maintenance matérielle (18%)         68 400 EUR/an
  Licences (NVIDIA AI Enterprise)      36 000 EUR/an
  Personnel MLOps (2.5 ETP)           225 000 EUR/an
  Monitoring + sécurité                15 000 EUR/an
  Total OPEX annuel                   369 400 EUR/an

TCO sur 3 ans
  CAPEX                               490 000 EUR
  OPEX (3 × 369 400)                1 108 200 EUR
  Total TCO 3 ans                   1 598 200 EUR
  Coût par token (50M tokens/jour) ≈ 0.000029 EUR
```



TCO Cloud : analyse par modèle de consommation

Le TCO cloud est fondamentalement différent dans sa structure : essentiellement composé d'OPEX, il élimine l'investissement initial mais génère des coûts récurrents proportionnels à l'usage. Pour un volume de **50 millions de tokens par jour** (volume typique d'une entreprise de taille intermédiaire avec 500 à 1 000 utilisateurs actifs de LLM), le coût mensuel varie considérablement selon le modèle choisi. Avec **GPT-4o** ($2.50/1M input, $10.00/1M output, ratio 2:1 input/output), le coût mensuel atteint environ 6 250 EUR pour l'input et 25 000 EUR pour l'output, soit 31 250 EUR/mois ou 375 000 EUR/an. Avec un modèle plus économique comme **Claude 3.5 Sonnet** ($3.00/$15.00 par million de tokens), le coût est comparable. L'utilisation de modèles plus petits comme **GPT-4o-mini** ($0.15/$0.60) réduit la facture à environ 1 875 EUR/mois, mais au prix d'une dégradation significative des performances sur les tâches complexes. À ces coûts d'API s'ajoutent les coûts d'infrastructure d'intégration (API gateway, load balancer, caching, logging), estimés à 2 000 à 5 000 EUR/mois, et le coût de personnel réduit mais non nul (1 à 1,5 ETP pour l'intégration et la maintenance des pipelines), soit 80 000 à 130 000 EUR/an.

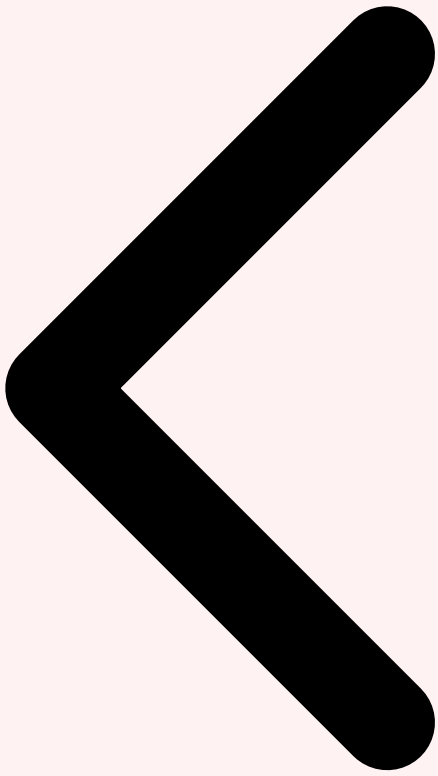


Le point de croisement : quand le on-premise devient rentable

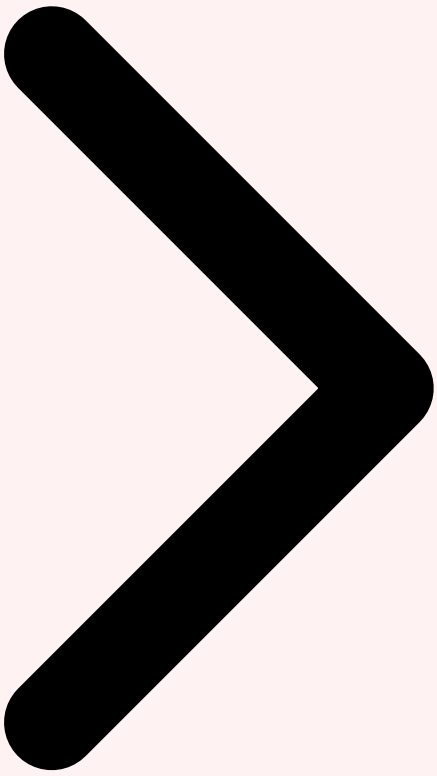
L'analyse comparative révèle un **point de croisement** (break-even point) au-delà duquel le déploiement on-premise devient plus économique que le cloud. Ce seuil dépend principalement du volume quotidien de tokens, du modèle cloud de référence et de la taille du cluster on-premise. Pour un cluster 8x H200 avec un modèle Llama 3.1 70B en FP8, le point de croisement se situe typiquement entre **15 et 30 millions de tokens par jour** par rapport à GPT-4o, et entre 50 et 100 millions de tokens par jour par rapport à GPT-4o-mini. En dessous de ce seuil, le cloud reste plus économique grâce à l'absence d'investissement initial et à l'élasticité native. Au-dessus, le on-premise génère des économies croissantes : à 100M tokens/jour, l'économie sur 3 ans atteint 500 000 à 1 200 000 EUR par rapport au cloud avec un modèle frontier. Ces chiffres supposent un taux d'utilisation GPU moyen de 60 à 75 %, ce qui est atteignable en production avec du continuous batching mais nécessite un flux de requêtes suffisamment régulier pour éviter les périodes de sous-utilisation qui dégradent la rentabilité.

- ▹**Coûts cachés cloud** — Egress data (transfert sortant), stockage des logs, backup, DDoS protection, WAF : ces postes ajoutent typiquement 10 à 20 % au coût API brut

- ▹**Coûts cachés on-premise** — Recrutement (3 à 6 mois pour un profil MLOps senior), formation continue, obsolescence matérielle (cycle GPU de 2 à 3 ans), coût d'opportunité du capital immobilisé
- ▹**Optimisation FinOps** — Les techniques de caching sémantique, de prompt compression et de routage intelligent (modèle léger pour les requêtes simples) peuvent réduire le coût cloud de 30 à 60 %
- ▹**Valeur résiduelle GPU** — Les GPU NVIDIA conservent 40 à 60 % de leur valeur après 3 ans sur le marché secondaire, améliorant significativement le TCO effectif on-premise

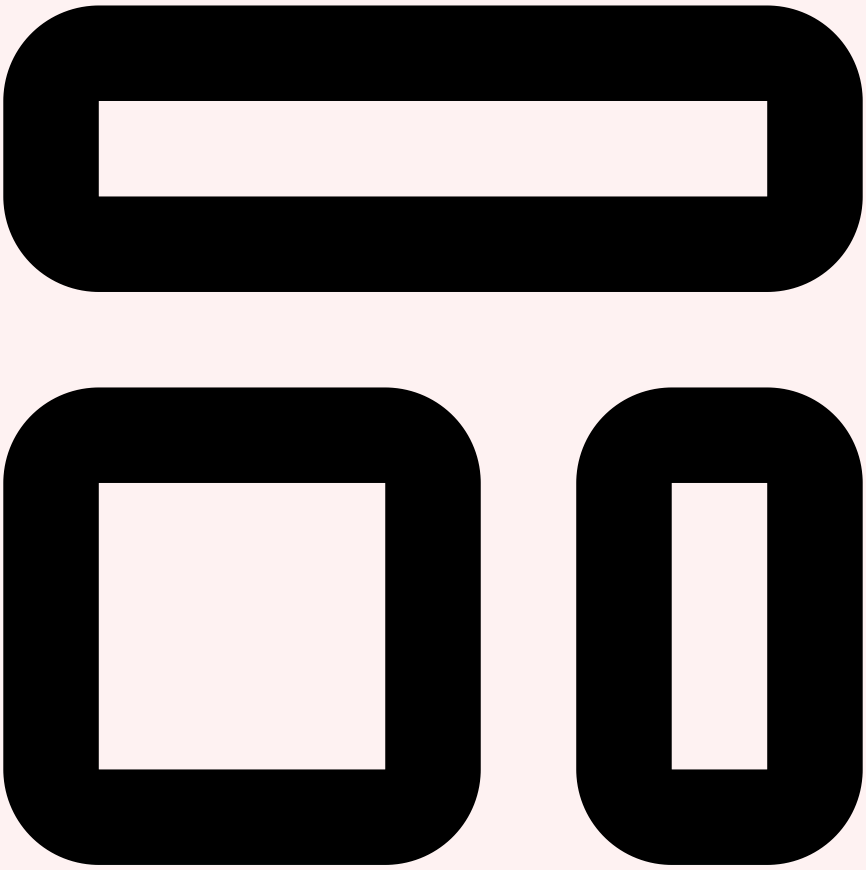


Souveraineté Comparatif Coûts Architectures Hybrides



6 Architectures Hybrides : Le Meilleur des Deux Mondes

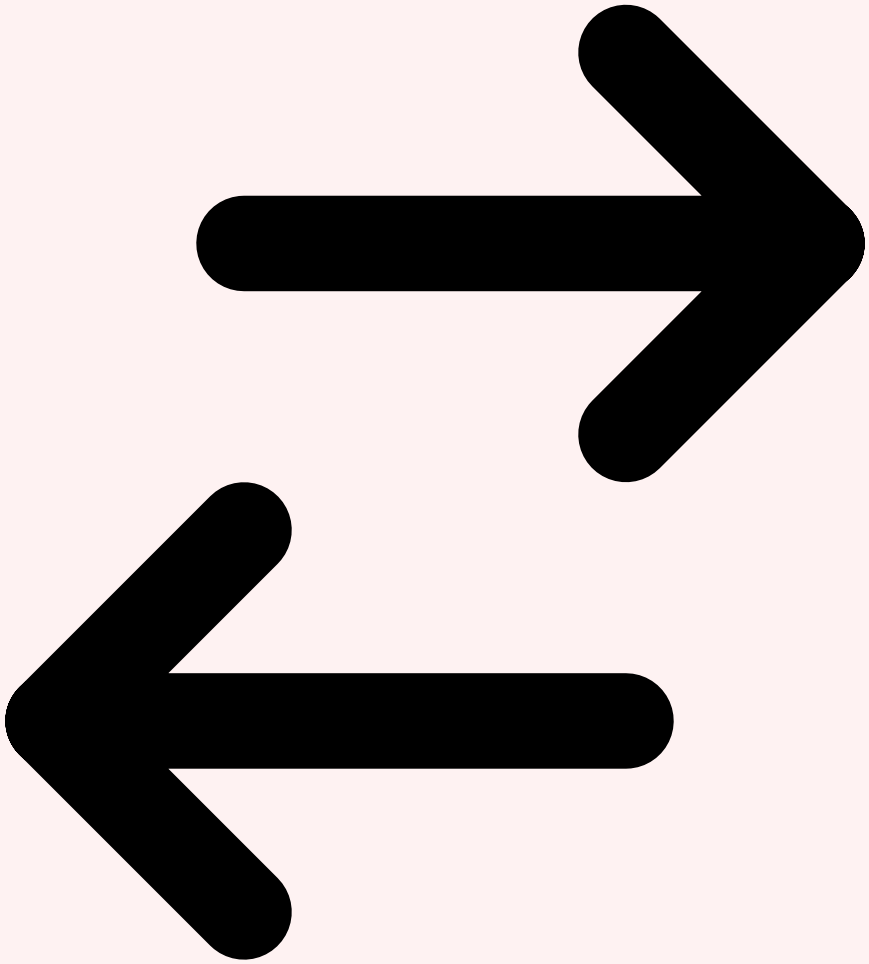
L'architecture hybride LLM s'impose en 2026 comme le choix pragmatique de la majorité des entreprises de taille intermédiaire et des grands groupes. Plutôt que de trancher de manière binaire entre cloud et on-premise, l'approche hybride **segmente les flux de données et les cas d'usage** pour diriger chaque requête vers l'infrastructure la plus adaptée en fonction de la sensibilité des données, de la complexité de la tâche et des exigences de latence. Cette architecture repose sur un **routeur intelligent** (LLM Router ou Gateway) qui analyse chaque requête entrante et la dirige vers le modèle et l'infrastructure optimaux. Les données sensibles — informations personnelles, données financières, secrets industriels, documents classifiés — sont systématiquement traitées par des modèles on-premise, tandis que les requêtes portant sur des données non sensibles ou publiques sont routées vers les APIs cloud pour bénéficier des modèles frontier les plus performants. Cette segmentation permet de **concilier souveraineté et performance** tout en optimisant le TCO global.



Architecture du LLM Router

Le composant central de l'architecture hybride est le **LLM Router**, un middleware intelligent qui intercepte chaque requête LLM et prend des décisions de routage en temps réel. L'implémentation de référence en 2026 repose sur plusieurs couches de décision. La première couche est un **classificateur de sensibilité** : un modèle léger (BERT ou DistilBERT fine-tuné) analyse le prompt en moins de 5 ms pour détecter la présence de données personnelles (NER), de données financières, de secrets commerciaux ou de tout contenu classifié selon la politique de l'organisation. La deuxième couche est un **estimateur de complexité** qui évalue si la requête nécessite un modèle frontier (raisonnement complexe, analyse juridique, code avancé) ou si un modèle plus léger suffit (résumé, traduction, Q&A factuel). La troisième couche applique les **règles de gouvernance** : quotas par département, budget maximum par utilisateur, blacklist de modèles pour certaines catégories de données. Le routage s'effectue en cascade : sensibilité d'abord, puis complexité, puis optimisation coût — la sécurité des données prime toujours sur les considérations économiques.

```yaml
# Architecture LLM Router — Configuration YAML
router:
  classification:
    model: distilbert-sensitivity-classifier-v3
    threshold: 0.85
    categories:
      - pii        # Données personnelles
      - financial   # Données financières
      - medical     # Données de santé
      - classified  # Documents confidentiels

  routes:
    sensitive:
      backend: on-premise
      models:
        - name: llama-3.1-70b-instruct
          endpoint: http://vllm-internal:8000/v1
          max_tokens: 8192
        - name: mistral-large-2-123b
          endpoint: http://vllm-internal:8001/v1
          max_tokens: 32768
      fallback: queue  # Jamais de fallback cloud

    standard:
      backend: cloud
      models:
        - name: gpt-4o
          provider: azure-openai
          region: francecentral
          priority: complex_tasks
        - name: claude-3.5-sonnet
          provider: aws-bedrock
          region: eu-west-3
          priority: analysis
      fallback: on-premise

  governance:
    budget_limits:
      daily_per_user: 50 EUR
      monthly_per_team: 5000 EUR
    audit:
      log_prompts: true
      log_responses: true
      retention: 90d
```

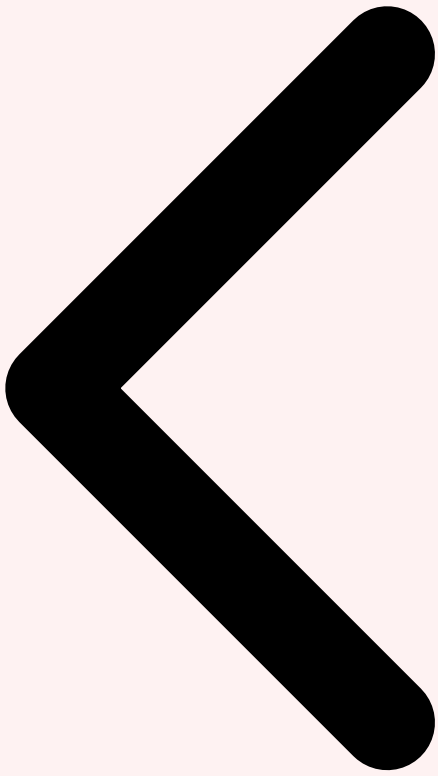


Patterns de déploiement hybride

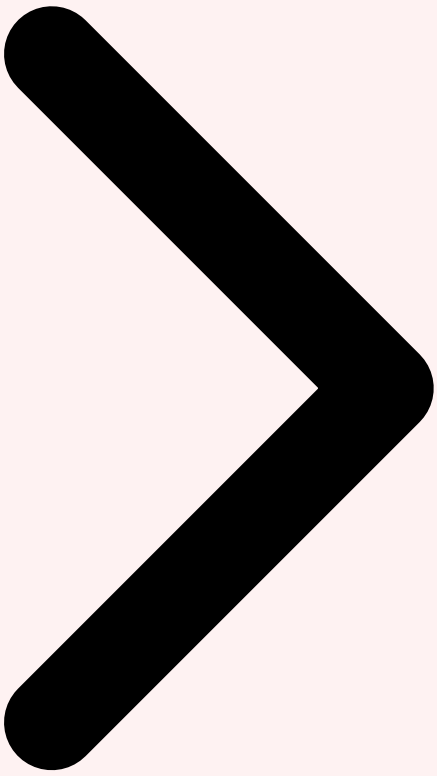
Trois patterns architecturaux dominent les déploiements hybrides en 2026. Le **pattern « Tiered Sensitivity »** est le plus courant : toutes les requêtes contenant des données sensibles sont traitées on-premise, le reste va au cloud. C'est le pattern recommandé pour les entreprises des secteurs régulés qui doivent démontrer la conformité RGPD et AI Act. Le **pattern « Cloud-First with On-Prem Fallback »** utilise le cloud comme backend principal pour maximiser les performances (accès aux modèles frontier), avec un fallback automatique vers les modèles on-premise en cas d'indisponibilité cloud, de dépassement de quotas ou de détection de données sensibles. Ce pattern convient aux entreprises qui privilégient la qualité des réponses et peuvent tolérer un risque résiduel maîtrisé. Le **pattern « On-Prem Primary with Cloud Burst »** utilise l'infrastructure on-premise comme backend principal pour les opérations courantes et bascule vers le cloud uniquement pour

absorber les pics de charge temporaires ou pour des tâches spécifiques nécessitant un modèle frontier non disponible localement. Ce pattern optimise le TCO pour les organisations avec un volume de base élevé et des pics prévisibles.

- ▹**DLP Gateway** — Déployer un Data Loss Prevention en amont du routeur pour anonymiser ou pseudonymiser automatiquement les données sensibles avant envoi cloud, avec ré-identification au retour
- ▹**Semantic Cache partagé** — Implémenter un cache sémantique (GPTCache, Redis avec embeddings) commun aux backends on-premise et cloud pour éviter les requêtes redondantes et réduire les coûts de 30 à 50 %
- ▹**Observabilité unifiée** — Centraliser les métriques de latence, coût, qualité et volume dans un dashboard unique (Grafana + Prometheus) couvrant les deux backends pour un pilotage FinOps efficace
- ▹**Failover automatique** — Configurer des circuit breakers avec des timeouts agressifs (3 à 5 secondes) et un basculement automatique vers le backend alternatif en cas de dégradation de service



Comparatif Coûts Architectures Hybrides Recommandations



7 Recommandations et Critères de Décision

Après avoir examiné en détail les trois modèles de déploiement, les enjeux de souveraineté, les comparatifs de coûts et les architectures hybrides, cette section synthétise les **recommandations opérationnelles** pour guider les décideurs techniques et stratégiques dans leur choix d'architecture LLM. Le choix optimal dépend d'une matrice de critères propre à chaque organisation, mais des patterns de décision clairs émergent de l'analyse des déploiements réussis en 2026. L'objectif n'est pas de désigner un modèle universellement supérieur — il n'existe pas — mais de fournir un **framework décisionnel structuré** qui intègre les dimensions technique, réglementaire, économique et organisationnelle, et de partager les retours d'expérience concrets qui permettent d'éviter les erreurs les plus courantes.

Matrice de décision par profil d'entreprise

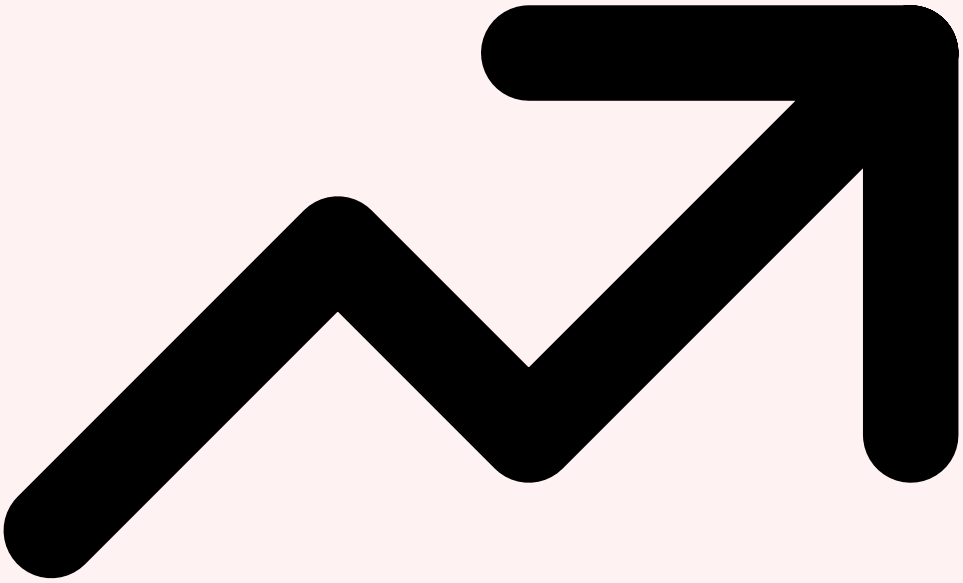
Les recommandations varient significativement selon le profil de l'organisation. Pour les **startups et PME innovantes** (moins de 500 collaborateurs, budget IA inférieur à 200K EUR/an), le cloud API est presque toujours le choix optimal : le time-to-market est immédiat, l'investissement initial est nul, et les volumes de tokens sont généralement insuffisants pour justifier un déploiement on-premise. La recommandation est d'utiliser des modèles cloud via des providers proposant des régions EU (Azure OpenAI en France Central, AWS Bedrock en eu-west-3) et d'implémenter un DLP minimaliste pour filtrer les données personnelles avant envoi. Pour les **ETI et grandes entreprises** (500 à 10 000 collaborateurs, budget IA de 500K à 5M EUR/an), l'architecture hybride s'impose comme le choix de référence : un cluster on-premise de 4 à 16 GPU pour les données sensibles et le fine-tuning, complété par des APIs cloud pour les cas d'usage non sensibles et les pics de charge. Pour les **grands groupes et organisations publiques** (plus de 10 000 collaborateurs, budget IA supérieur à 5M EUR/an, contraintes réglementaires fortes), le on-premise constitue souvent le coeur de la stratégie, avec un cloud souverain qualifié SecNumCloud comme extension pour l'élasticité.

Critère	Privilégier Cloud	Privilégier On-Premise	Privilégier Hybride
Volume tokens	< 10M/jour	> 50M/jour	10-50M/jour
Sensibilité données	Publiques / internes	Confidentielles / secrètes	Mix sensible / non-sensible
Budget CAPEX	< 200K EUR	> 500K EUR	300-800K EUR
Équipe MLOps	0-1 ETP	3+ ETP	2-3 ETP
Latence requise	> 200ms acceptable	< 20ms critique	Variable par use case
Réglementation	Faible / standard	OIV / santé / défense	Sectoriel (banque, assurance)
Time-to-market	Critique (< 1 mois)	Planifié (3-6 mois)	Progressif (2-4 mois)

Les erreurs à éviter en 2026

L'analyse des déploiements LLM échoués ou sous-optimaux en 2025-2026 révèle des patterns d'erreur récurrents. La première erreur est le **« surdimensionnement initial »** : investir dans un cluster de 32 GPU avant d'avoir validé les cas d'usage en production. La recommandation est de commencer avec un cluster minimal (4 à 8 GPU) et de scaler

progressivement en fonction de la demande réelle. La deuxième erreur est la **« sous-estimation du coût humain »** : les GPU ne s'administrent pas tout seuls, et le recrutement d'ingénieurs MLOps compétents en infrastructure GPU est un processus qui prend 3 à 6 mois en 2026, avec des salaires en forte hausse. La troisième erreur est le **« tout-cloud sans gouvernance »** : déployer des APIs LLM en libre-service sans contrôle des coûts, des données envoyées et de la qualité des réponses, ce qui mène à du shadow AI, des dépassements budgétaires et des risques de conformité. La quatrième erreur est l'**« optimisation prématurée du TCO »** : passer 6 mois à construire une infrastructure on-premise avant de valider que le cas d'usage génère de la valeur métier, alors qu'un prototype cloud aurait permis de valider le ROI en 2 semaines. Pour approfondir, consultez Evasion d'EDR/XDR : techniques.



Roadmap d'implémentation recommandée

La roadmap que nous recommandons pour les entreprises qui abordent le déploiement LLM en production suit une progression en quatre phases. **Phase 1 (Mois 1-2) : Validation cloud** — Déployer les premiers cas d'usage via APIs cloud avec un DLP basique, mesurer les volumes, la qualité et le ROI. **Phase 2 (Mois 3-4) : Gouvernance et classification** — Implémenter la classification des données, le LLM Router, les politiques de sécurité et le

framework d'évaluation de la qualité. **Phase 3 (Mois 5-8) : Infrastructure on-premise** — Déployer le cluster GPU initial, migrer les workloads sensibles et à haut volume vers le on-premise, valider les performances et la conformité. **Phase 4 (Mois 9-12) : Optimisation hybride** — Affiner le routage, optimiser le TCO, déployer le semantic cache, implémenter le FinOps et automatiser le scaling. Cette approche progressive minimise le risque, valide la valeur métier avant l'investissement lourd, et permet à l'équipe de monter en compétence graduellement sur les technologies GPU et MLOps.

Conclusion : Le choix entre cloud, on-premise et hybride pour le déploiement de LLM n'est pas une décision binaire mais un **continuum architectural** que chaque organisation doit positionner en fonction de ses contraintes propres. La tendance dominante en 2026 est clairement à l'**architecture hybride**, qui permet de concilier souveraineté, performance et maîtrise des coûts. Les organisations qui réussiront le mieux leur transformation IA seront celles qui auront su construire une **infrastructure flexible et gouvernée**, capable d'évoluer au rythme de l'innovation technologique (nouveaux modèles, nouveaux GPU, nouvelles réglementations) tout en maintenant un contrôle rigoureux sur la sécurité des données et la conformité réglementaire. L'essentiel est de **commencer vite, itérer progressivement et gouverner rigoureusement**.

- ▹**Éviter le lock-in** — Utiliser l'API compatible OpenAI pour tous les backends (vLLM, TGI, cloud) afin de pouvoir migrer les workloads entre on-premise et cloud sans refactoring
- ▹**Investir dans l'évaluation** — Mettre en place un framework d'évaluation automatisé (LLM-as-judge, benchmarks métier, A/B testing) pour comparer objectivement les modèles cloud et on-premise sur vos cas d'usage réels
- ▹**Planifier la scalabilité** — Dimensionner l'infrastructure réseau et le datacenter pour supporter un doublement de la capacité GPU à 12-18 mois, le volume de requêtes LLM croissant en moyenne de 100 % par an
- ▹**Documenter la conformité** — Maintenir un registre de traitements IA (exigence AI Act) qui documente chaque système LLM déployé, sa base légale, sa classification de risque et les mesures de mitigation associées

Besoin d'un accompagnement expert ?

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

Références et ressources externes

- vLLM — Moteur d'inférence LLM haute performance
- llama.cpp — Inférence LLM optimisée en C/C++
- MLflow — Plateforme open source de gestion du cycle de vie ML
- Kubernetes Docs — Documentation officielle Kubernetes
- 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 LLM On-Premise vs Cloud ?

Le concept de LLM On-Premise vs Cloud 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 LLM On-Premise vs Cloud est-il important en cybersécurité ?

La compréhension de LLM On-Premise vs Cloud permet aux équipes de sécurité d'améliorer leur posture défensive. Les sections « Table des Matières » et « 1 Les Enjeux du Déploiement de LLM en 2026 » 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 Enjeux du Déploiement de LLM en 2026, 2 Déploiement Cloud : APIs et Managed Services. 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