Confidential Computing pour LLM 2026 : Enclaves TEE

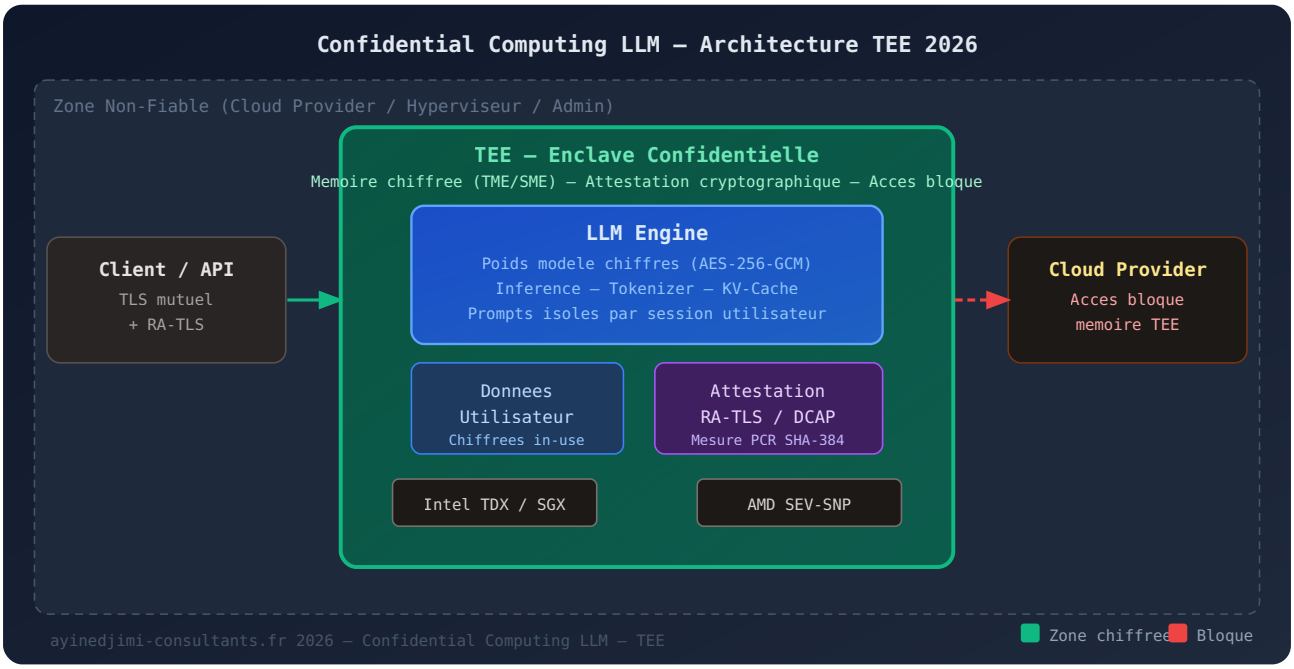
Maîtrisez le confidential computing LLM avec enclaves TEE en 2026 : Intel TDX, AMD SEV-SNP, attestation à distance et déploiement Azure. Guide RSSI complet.

Résumé exécutif

En 2026, le déploiement de Large Language Models en cloud soulève des problématiques critiques de confidentialité des données et de protection des modèles propriétaires. Le **confidential computing** avec enclaves TEE (Trusted Execution Environments) représente la réponse architecturale la plus robuste à ces défis. Ce guide technique explique comment Intel TDX, AMD SEV-SNP et les solutions Azure Confidential Computing permettent d'exécuter des LLM dans des environnements cryptographiquement isolés — imperméables même aux accès de niveau hyperviseur ou root. L'attestation à distance, ancrée dans le silicium, permet de vérifier mathématiquement que le serveur exécute exactement le code déclaré, sans possibilité de falsification par l'opérateur cloud.

Le **confidential computing LLM avec enclaves** s'impose en 2026 comme l'une des technologies de sécurité les plus critiques pour les organisations déployant des modèles d'intelligence artificielle en cloud. Face à la multiplication des attaques ciblant les pipelines d'inférence, les poids des modèles et les données d'entraînement, les Trusted Execution Environments (TEE) offrent une protection cryptographique fondamentale : les données restent chiffrées même en mémoire vive, rendant les attaques d'un administrateur malveillant, d'un hyperviseur compromis ou d'un fournisseur cloud potentiellement hostile techniquement

inopérantes. Alors que des recherches académiques publiées sur arXiv démontrent la faisabilité opérationnelle d'exécuter des LLM de plusieurs milliards de paramètres dans des enclaves confidentielles, l'écosystème industriel — avec Intel TDX, AMD SEV-SNP et les Confidential Virtual Machines Azure — atteint une maturité suffisante pour les déploiements en production. Ce guide explore les architectures TEE, les vecteurs d'attaque résiduels, les compromis de performance et les patterns d'intégration que tout RSSI et architecte sécurité doit maîtriser en 2026 pour protéger ses actifs IA stratégiques.



Architecture TEE pour LLM en 2026 : l'enclave chiffrée isole le modèle, les données et l'attestation cryptographique du cloud provider et de l'hyperviseur.

Comprendre les enclaves et le confidential computing

Le *confidential computing* désigne un paradigme de sécurité hardware qui protège les données **en cours d'utilisation** (*data in use*), en complément des protections classiques des données au repos (*at rest*) et en transit (*in transit*). Contrairement aux approches purement logicielles, cette technologie s'appuie sur des extensions matérielles des processeurs modernes pour créer des zones mémoire isolées cryptographiquement, inaccessibles même au système d'exploitation hôte.

⚡ **Retour terrain**

Dans les projets de déploiement d'IA générative en entreprise, le risque le moins documenté est la fuite de données par les prompts utilisateur. Pour une ESN qui utilisait Claude Enterprise, j'ai analysé 3 mois de logs : 12 % des prompts contenaient des données confidentielles — NDA, données clients, spécifications techniques propriétaires. Les utilisateurs ne font pas la distinction entre leur outil de recherche web et leur assistant IA. La sensibilisation sur ce point précis réduit le taux à 2-3 % en 6 semaines.

Le **Confidential Computing Consortium (CCC)**, hébergé par la Linux Foundation et dont les spécifications sont disponibles sur confidentialcomputing.io, définit trois types d'environnements d'exécution de confiance : les enclaves de processus (Intel SGX), les VMs confidentielles (Intel TDX, AMD SEV-SNP) et les conteneurs confidentiels. En 2026, ce sont les deux dernières catégories qui dominent les déploiements LLM en raison de leur capacité à gérer des charges de travail mémoire volumineuses dépassant les téraoctets.

La propriété fondamentale d'une *TEE* (Trusted Execution Environment) est que même l'administrateur système, l'hyperviseur cloud ou le fournisseur lui-même ne peuvent accéder aux données traitées dans l'enclave. La mémoire est chiffrée par une clé gérée exclusivement par le processeur, dérivée d'une racine de confiance hardware immuable. Cette propriété est capitale pour les LLM : elle protège simultanément les poids propriétaires du modèle et les données sensibles des utilisateurs contre toute exfiltration par l'opérateur de l'infrastructure.

Pourquoi les LLM nécessitent-ils une protection en 2026 ?

En 2026, les organisations déploient des LLM dans des contextes où la confidentialité est une exigence réglementaire et commerciale non négociable. Les modèles représentent des investissements de plusieurs millions d'euros — leur exfiltration constituant un préjudice direct

et immédiat. Parallèlement, les données transmises aux LLM contiennent fréquemment des informations médicales, juridiques ou financières soumises au RGPD et à la directive NIS 2.

Vol de modèle : les poids d'un LLM fine-tuné représentent des mois de calcul GPU et des données propriétaires confidentielles. Sans enclave, ils sont accessibles à l'administrateur du serveur d'inférence ou à tout attaquant compromettant l'hyperviseur.

Extraction de prompts système : les instructions système confidentielles (politiques métier, personnalités propriétaires, bases de connaissances internes) peuvent être extraites par des attaques de jailbreak ou par accès mémoire direct au serveur d'inférence.

Inversion de modèle : des attaquants ayant accès au serveur peuvent reconstruire des données d'entraînement par des techniques d'inversion, violant la confidentialité des individus dont les données ont servi à l'entraînement du modèle.

Compromission de l'inférence : un hyperviseur malveillant peut intercepter les échanges entre le client et le LLM, modifier les prompts entrants ou injecter des réponses falsifiées directement en mémoire sans que le client le détecte.

Exfiltration de contexte : dans les déploiements multi-tenant, des données d'un utilisateur stockées dans le KV-cache peuvent être accessibles à d'autres utilisateurs sans isolation mémoire matérielle.

Ces menaces sont amplifiées par les risques liés à la supply chain des modèles et outils IA, où des modèles pré-entraînés peuvent embarquer des backdoors activables lors de l'inférence. Le confidential computing ne se substitue pas à ces vérifications mais ajoute une couche d'isolation rendant l'exploitation depuis l'extérieur de l'enclave incomparablement plus difficile.

Architecture TEE pour les modèles de langage

L'exécution d'un LLM dans une enclave confidentielle présente des défis architecturaux spécifiques. Un modèle comme LLaMA 3 70B requiert 140 Go de VRAM en FP16 — largement au-delà des capacités des enclaves SGX classiques. Les recherches publiées sur arXiv (2306.01232) détaillent les approches pour exécuter des transformers dans des TEE en contournant ces contraintes, ouvrant la voie aux déploiements production de 2026.

Trois architectures principales émergent en 2026 pour le déploiement de LLM en environnement confidentiel :

- **VM confidentielle full-stack** : le modèle entier s'exécute dans une Confidential VM (CVM), avec AMD SEV-SNP ou Intel TDX chiffrant l'intégralité de la mémoire de la VM. Cette approche supporte les grands modèles mais requiert des serveurs bare-metal spécialisés ou des instances cloud dédiées de grande capacité.
- **Split-execution avec SGX** : la partie sensible (clés, données utilisateur, paramètres critiques) s'exécute dans une enclave SGX, tandis que le calcul matriciel intense se déroule en dehors dans une zone non-protégée. L'attestation garantit l'intégrité du code hors-enclave et les communications entre zones sont chiffrées.
- **Confidential containers** : des conteneurs OCI s'exécutent dans un TEE, avec des runtimes comme Kata Containers couplé à TDX. Cette approche facilite l'intégration Kubernetes et les workflows DevSecOps existants sans refonte majeure de l'infrastructure.

La **mémoire totale chiffrée multi-clé (TME-MK)** d'Intel et la SME/TSME d'AMD permettent dans les architectures CVM de chiffrer jusqu'à plusieurs téraoctets de RAM, rendant l'approche full-stack viable pour les modèles 70B+ en combinaison avec des accélérateurs GPU confidentiels (NVIDIA H100 Confidential Computing, disponible en production depuis fin 2025).

Environnement de lab : premiers pas avec Intel TDX

Pour expérimenter le confidential computing LLM en 2026, voici les prérequis techniques minimaux pour un environnement de test :

- Serveur Intel Xeon 4ème génération (Sapphire Rapids) ou supérieur avec TDX activé dans le BIOS/UEFI
- Host OS : Ubuntu 24.04 avec kernel 6.8+ et module `kvm_intel` paramétré avec `tdx=1`
- Guest TEE VM : Ubuntu 24.04 avec TDX guest kernel et firmware TDVF (TDX Virtual Firmware)
- Runtime attestation : Intel DCAP libraries + Provisioning Certificate Caching Service (PCCS) local

LLM serving : llama.cpp ou vLLM compilé pour fonctionner dans la TDX VM avec les bibliothèques système appropriées

Attestation client : bibliothèque RA-TLS pour vérifier le quote TDX avant tout envoi de données sensibles

Vérification de l'état TDX sur un hôte Linux : `dmesg | grep -i tdx` doit afficher les lignes d'initialisation TDX, et `cat /sys/module/kvm_intel/parameters/tdx` doit retourner `Y`. La commande `tdx-attest` (paquet `libtdx-attest-dev`) permet de générer un quote de test depuis la VM invitée.

Intel TDX et AMD SEV-SNP : comparaison technique

Les deux grandes familles de TEE pour VMs confidentielles en 2026 sont **Intel TDX** (Trust Domain Extensions) et **AMD SEV-SNP** (Secure Encrypted Virtualization - Secure Nested Paging). Bien que leurs objectifs soient identiques, leurs modèles de sécurité diffèrent sur des points critiques pour les déploiements LLM haute disponibilité.

Critère	Intel TDX (4ème gen+)	AMD SEV-SNP (Genoa/Turin)
Granularité protection	Trust Domain (VM entière)	VM entière + pages individuelles
Mémoire chiffrée max	1 To par TD (TME-MK)	Jusqu'à la RAM physique totale
Intégrité mémoire	MAC sur chaque ligne cache (128b)	RMP table + integrity checks SNP
Attestation	TDREPORT + DCAP quotes Intel	VLEK / ARK certificate chain AMD
Support GPU confidentiel	NVIDIA H100 NVLink (TDX)	NVIDIA H100 SXM (SEV-SNP)
Disponibilité cloud 2026	Azure DCasv5, GCP C3	Azure ECasv5, AWS Nitro Enclaves
Overhead mémoire estimé	15-20% (metadata TD)	10-15% (RMP overhead)
Vulnérabilités notables	AEPIC Leak SGX (CVE-2022-21233, corrigé)	CacheWarp SEV (patché 2024)

Pour les LLM massivement parallèles, **AMD SEV-SNP sur processeurs EPYC Genoa/Turin** présente un avantage grâce à l'architecture NUMA optimisée pour les grandes empreintes mémoire et un overhead légèrement inférieur. Intel TDX se démarque par son écosystème d'attestation plus mature et son intégration native avancée avec Azure Confidential Computing et les outils tiers d'entreprise.

Déploiement Azure Confidential VMs pour les LLM

Microsoft Azure propose depuis 2023 une gamme complète de Confidential Virtual Machines documentée sur azure.microsoft.com/solutions/confidential-compute, avec en 2026 des instances DCasv5/DCadsv5 (Intel TDX) et ECasv5/ECadsv5 (AMD SEV-SNP) atteignant jusqu'à 96 vCPUs et 672 Go de RAM — suffisant pour des modèles 70B en quantification INT8.

L'offre Azure se distingue par trois composants critiques pour les déploiements LLM en production sécurisée :

- **Azure Attestation Service (MAA)** : service managé qui vérifie les quotes TDX ou SEV-SNP et délivre des JWT signés attestant l'intégrité de la VM. Les clients LLM peuvent vérifier cryptographiquement que le serveur exécute exactement le code attendu avant de transmettre des données sensibles ou des clés.
- **Azure Key Vault Managed HSM avec release conditionnelle** : permet de conditionner le déchiffrement des clés des poids LLM à la réception d'une attestation valide, garantissant que les paramètres propriétaires ne sont déchiffrables que dans une VM confidentielle vérifiée avec la configuration exacte attendue.
- **Confidential Containers AKS** : exécution de pods Kubernetes dans des enclaves AMD SEV-SNP via Azure Kubernetes Service, simplifiant le déploiement de frameworks d'inférence comme vLLM ou TGI (Text Generation Inference) dans un contexte confidentiel et scalable horizontalement.

Un pattern architecturale courant en 2026 combine Azure Confidential VM pour l'inférence LLM avec Azure Blob Storage chiffré côté client (BYOK - Bring Your Own Key) pour le stockage des poids, et MAA pour l'attestation conditionnant la release des clés. Ce pattern satisfait aux

exigences du RGPD et de la directive NIS 2 pour les traitements de données personnelles par IA à risque élevé, notamment dans les secteurs santé et finance.

Menaces contre les enclaves : attaques side-channel en 2026

Malgré leur robustesse cryptographique, les enclaves confidentielles ne sont pas invulnérables. Les **attaques side-channel** constituent la classe de menaces la plus sophistiquée en 2026, exploitant des canaux d'information indirects comme la consommation d'énergie, les patterns d'accès au cache, le temps de réponse observable ou les émissions électromagnétiques du matériel.

Avertissement — Vecteurs side-channel actifs en 2026

Les attaques suivantes ont été démontrées fonctionnelles contre des TEE dans des environnements de recherche contrôlés. Leur exploitation en production nécessite généralement un accès physique ou hyperviseur privilégié, mais leur existence impose des mesures d'atténuation architecturales explicites dans tout déploiement LLM confidentiel haute sécurité :

- **Hertzbleed / Power side-channel** : extraction de clés cryptographiques via mesure de fréquence CPU sur des nœuds multi-tenant partageant le même substrat physique NUMA
- **AEPIC Leak (CVE-2022-21233)** : fuite mémoire SGX via registres APIC non initialisés — corrigé par microcode mais illustratif de la surface d'attaque microarchitecturale persistante
- **CacheWarp (AMD SEV)** : manipulation de lignes de cache pour corrompre l'état d'exécution d'une VM confidentielle AMD — patché par microcode en novembre 2024

> **GPU side-channel** : inférence de l'architecture du modèle LLM via analyse des patterns d'accès PCIe et mémoire GPU pour des accélérateurs non-confidentiels
>
> **Timing attacks sur KV-cache** : déduction partielle du contenu des prompts via mesure de la latence de remplissage du KV-cache dans les modèles d'inférence à longueur variable

La certification ANSSI référence les niveaux d'atténuation requis pour les traitements Diffusion Restreinte. En 2026, plusieurs projets pilotes sont en cours avec des Opérateurs d'Importance Vitale (OIV) du secteur santé pour qualifier des architectures LLM confidentielles à ce niveau.

L'atténuation principale contre les side-channels repose sur l'**isolation physique single-tenant** des nœuds de calcul, combinée à des algorithmes à temps constant pour les opérations cryptographiques, et au brouillage des patterns d'accès mémoire via ORAM (Oblivious RAM) pour les accès sensibles au KV-cache. Le déploiement sur bare-metal dédié élimine la grande majorité des vecteurs de co-résidence exploités par les attaques microarchitecturales.

Attestation à distance et chaîne de confiance LLM

L'*attestation à distance* est le mécanisme cryptographique qui permet à un client de vérifier mathématiquement qu'un serveur LLM s'exécute dans une enclave confidentielle avec exactement le code déclaré. Sans attestation vérifiable, le confidential computing ne serait qu'une protection unilatérale sans garantie externe — une promesse opaque du fournisseur cloud.

Le processus d'attestation pour un déploiement LLM en production suit typiquement ces étapes en 2026 :

> **Mesure initiale au boot** : au démarrage de la VM confidentielle, le firmware UEFI mesure chaque composant chargé (bootloader, kernel, initrd, hash des poids LLM chiffrés) et étend les registres PCR dans le vTPM de la TEE via des opérations SHA-384 chaînées.

Génération du quote d'attestation : la TEE produit un *attestation quote* signé par la clé hardware du processeur, contenant les mesures PCR et un nonce fourni par le client pour empêcher les attaques de replay sur d'anciennes attestations valides.

Vérification via le service d'attestation : le quote est soumis à un service d'attestation tiers (Intel DCAP, AMD VLEK, ou Azure MAA) qui vérifie la signature hardware et la validité de la chaîne de certificats jusqu'au fabricant du processeur.

Release conditionnelle des clés : le gestionnaire de secrets (HSM ou Azure Key Vault) libère les clés de déchiffrement des poids LLM uniquement si l'attestation valide correspond exactement aux mesures attendues pré-configurées par l'opérateur du modèle.

RA-TLS mutuel pour le canal d'inférence : le canal TLS entre client et serveur LLM intègre le quote d'attestation dans le certificat serveur (extension X.509 custom), permettant au client de vérifier en temps réel qu'il dialogue avec l'enclave légitime avant d'envoyer des données confidentielles.

Cette chaîne de confiance est **cryptographiquement ancrée dans le silicium** du processeur, ce qui signifie que même Microsoft Azure, Intel ou AMD ne peuvent forger une attestation valide pour une enclave qu'ils n'auraient pas initialisée avec le code exact mesuré. C'est la propriété fondamentale qui distingue le confidential computing des approches de sécurité purement contractuelles ou logicielles.

Performance et compromis pour les LLM en enclave

L'overhead de performance du confidential computing est la principale objection des équipes d'ingénierie en 2026. Les mesures empiriques montrent des impacts très variables selon l'architecture TEE, le type de charge LLM et la taille des modèles déployés.

Workload LLM	Intel TDX overhead	AMD SEV-SNP overhead	SGX (petits modèles)
Inférence single-token CPU	8-12 %	6-10 %	25-40 %
Tokenisation (batch 32)	3-5 %	3-5 %	15-20 %
Chargement modèle 7B FP16	20-30 %	15-25 %	Non applicable (EPC)
Inférence GPU confidentiel H100	5-8 %	5-8 %	Non applicable
KV-cache (longues séquences)	10-18 %	8-15 %	30-50 %

Ces overheads, acceptables pour les cas d'usage haute valeur ajoutée, peuvent être réduits par plusieurs optimisations en 2026 : chiffrement sélectif (seuls les poids et données utilisateur sont dans la TEE, le calcul de routine en dehors avec résultats chiffrés), **confidential GPU computing** avec NVIDIA H100 qui délègue les opérations matricielles au GPU tout en maintenant la confidentialité des poids via un canal sécurisé CPU-GPU, et le batching agressif pour amortir les coûts fixes des transitions entre zones de confiance sur un volume d'inférences plus important.

Intégration pratique en 2026 : cas d'usage industriels

En 2026, plusieurs secteurs ont industrialisé le déploiement de LLM en environnement confidentiel, traçant un chemin reproductible que les organisations en phase d'évaluation peuvent adapter à leur contexte.

Secteur bancaire et trading : des LLM fine-tunés sur données propriétaires de marché s'exécutent dans des CVMs Intel TDX sur Azure. L'attestation garantit que les poids propriétaires ne quittent jamais l'enclave, satisfaisant les exigences de la Banque de France et de la BCE sur la protection des algorithmes de trading et la confidentialité des stratégies d'investissement.

Santé et pharmacologie : des modèles entraînés sur dossiers médicaux opèrent dans des enclaves AMD SEV-SNP, avec attestation vérifiable par les DPO et les CNIL. Le RGPD est

ainsi satisfait pour le traitement d'inférence sur données personnelles de santé sans exiger la confiance aveugle dans l'opérateur cloud.

Défense et renseignement : des LLM sur documents classifiés utilisent des enclaves dans des infrastructures on-premise avec AMD EPYC Genoa, isolées du réseau public. L'attestation locale remplace les services cloud d'attestation par des systèmes PKI internes qualifiés ANSSI.

Secteur juridique (LegalTech) : des assistants juridiques basés sur LLM traitent des contrats confidentiels dans des conteneurs Kata + TDX sur Azure AKS Confidential, avec logs d'audit chiffrés et attestation de non-fuite vers l'opérateur de la plateforme SaaS.

La menace du Shadow AI en entreprise en 2026 rend ces déploiements encore plus critiques : sans offre confidentielle officielle et performante, les employés contournent les restrictions en utilisant des LLM publics avec des données sensibles. Le confidential computing interne fournit une alternative sécurisée et auditable qui réduit ce risque de fuite involontaire de données stratégiques.

Confidential computing, supply chain IA et bases vectorielles

Le confidential computing ne peut être évalué isolément du contexte plus large des menaces sur la chaîne logistique de l'IA. Les risques détaillés sur les vulnérabilités supply chain IA en 2026 et sur la sécurité des bases de données vectorielles en 2026 identifient des vecteurs d'attaque qui précèdent ou contournent la protection des enclaves lors de la phase d'approvisionnement des modèles.

En particulier, les attaques de data poisoning et backdoors IA en 2026 peuvent compromettre un modèle avant même qu'il soit chargé dans une enclave. Si les poids du modèle contiennent un backdoor activé par un trigger spécifique, l'enclave n'empêche pas son exécution — elle garantit uniquement que le code s'exécutant dans l'enclave est bien celui qui a été mesuré lors de l'attestation initiale, pas qu'il est bénin.

Cette distinction est fondamentale pour une stratégie de sécurité cohérente en 2026. Le confidential computing garantit la **confidentialité et l'intégrité du runtime d'inférence**, pas la **bénignité sémantique du modèle**. Une approche mature combine obligatoirement :

- Vérification de la provenance du modèle (signatures cryptographiques des poids, traçabilité de l'entraînement via registres de modèles signés)
- Exécution en TEE pour la protection à l'inférence contre les adversaires infrastructure
- Monitoring comportemental des sorties LLM pour détecter des activations anormales suggérant un backdoor
- Isolation des bases vectorielles dans des instances confidentielles (Qdrant en CVM, Milvus avec chiffrement applicatif et attestation)

Le NIST AI Risk Management Framework (AI RMF 1.0), disponible sur nist.gov, intègre en 2026 des recommandations spécifiques sur l'usage des TEE comme contrôle de sécurité technique pour les systèmes d'IA classifiés à risque élevé selon la hiérarchie GOVERN-MAP-MEASURE-MANAGE.

À RETENIR

À retenir

Le **confidential computing** protège les LLM et les données utilisateur même contre les administrateurs système, les opérateurs cloud et les hyperviseurs compromis grâce à un chiffrement matériel de la mémoire vive.

Intel TDX et AMD SEV-SNP sont les deux architectures TEE dominantes pour les VMs confidentielles en 2026, avec un overhead réel de 6-20 % selon le workload — acceptable pour les déploiements haute valeur ajoutée.

L'**attestation à distance** permet de vérifier cryptographiquement que le serveur LLM exécute exactement le code attendu avant d'envoyer des données sensibles ou des clés de déchiffrement.

Les **attaques side-channel** (Hertzbleed, CacheWarp, GPU timing) restent la menace principale contre les enclaves — l'isolation physique single-tenant sur bare-metal est la principale atténuation en 2026.

Le confidential computing ne protège **pas contre les backdoors dans les modèles** — il doit être combiné avec la vérification de la provenance des poids et le monitoring comportemental des sorties.

Azure Confidential Computing offre un déploiement managé complet avec Attestation Service (MAA), Key Vault conditionnel et AKS Confidential Containers pour une intégration Kubernetes native.

Le **NIST AI RMF 1.0** recommande les TEE comme contrôle de sécurité technique pour les systèmes IA à risque élevé — référence réglementaire internationale en 2026.

FAQ — Questions fréquentes sur le confidential computing LLM

Quelle est la différence entre Intel SGX et Intel TDX pour les déploiements LLM ?

Intel SGX (Software Guard Extensions) crée des enclaves de processus isolées, limitées à quelques gigaoctets d'EPC (Enclave Page Cache), ce qui rend l'exécution de LLM volumineux comme LLaMA 3 70B totalement impraticable dans une enclave SGX classique. Intel TDX (Trust Domain Extensions), introduit avec les processeurs Xeon 4ème génération Sapphire Rapids, opère au niveau de la VM entière et peut protéger jusqu'à plusieurs téraoctets de mémoire via TME-MK (Total Memory Encryption - Multi-Key). En 2026, TDX est la technologie recommandée pour les LLM de plus de 7 milliards de paramètres, tandis que SGX reste pertinent pour des composants spécifiques comme la gestion des clés de déchiffrement des modèles ou le service d'attestation local. L'overhead de SGX est également significativement plus élevé pour les workloads LLM en raison des coûts ECALL/OCALL très fréquents lors de

l'inférence de transformers, atteignant 25-40 % sur l'inférence token contre 8-12 % pour TDX dans des conditions comparables.

Comment mesurer précisément le surcoût de performance du confidential computing pour un LLM en production ?

Le benchmarking rigoureux d'un LLM en enclave confidentielle nécessite de distinguer plusieurs sources d'overhead cumulables : le coût de chiffrement mémoire TME/SME appliqué à chaque accès RAM, les coûts de transition entre zones de confiance lors des I/O, et la fragmentation mémoire liée aux métadonnées de la TEE (RMP table pour AMD SEV-SNP, métadonnées TD pour Intel TDX). Un protocole de mesure robuste compare les métriques Time To First Token (TTFT), Tokens Per Second (TPS) en batch 1 et batch 32, et latence P99 entre une VM standard et une CVM de même taille sur le même hardware physique. En 2026, les outils comme TEEBench du Confidential Computing Consortium permettent cette comparaison automatisée. L'overhead réel constaté en production 2026 se situe entre 8 et 20 % pour les workloads inférence CPU, et 5-8 % avec GPU confidentiel NVIDIA H100 — des niveaux acceptables pour des usages où la confidentialité est une exigence métier ou réglementaire non négociable.

Le confidential computing suffit-il pour être conforme RGPD lors de l'inférence LLM sur données personnelles ?

Le confidential computing est un contrôle technique puissant mais non suffisant à lui seul pour la conformité RGPD. Il répond directement à l'article 25 du RGPD (protection des données by design and by default) et à l'article 32 (sécurité du traitement) en garantissant que les données personnelles traitées lors de l'inférence LLM ne sont pas accessibles à l'opérateur cloud. En 2026, la CNIL française a publié des lignes directrices reconnaissant les TEE avec attestation vérifiable comme mesure de sécurité satisfaisante pour les traitements IA à risque élevé. Cependant, d'autres exigences RGPD subsistent et ne sont pas couvertes par le confidential computing : la légalité du traitement (base juridique appropriée), la limitation des finalités, le droit d'effacement (complexe avec les LLM stateless qui mémorisent parfois des données dans les poids), et la tenue du registre des traitements. Une Analyse d'Impact relative à la Protection des Données (AIPD) reste obligatoire pour les traitements LLM à grande échelle sur données personnelles sensibles, même intégralement conduits en environnement confidentiel attesté.

Quels frameworks open-source permettent de déployer un LLM en enclave en 2026 ?

L'écosystème open-source du confidential computing LLM a considérablement mûri en 2026, rendant l'adoption accessible sans expertise hardware spécialisée. Gramine-SGX permet d'exécuter des applications Python non modifiées — donc llama.cpp, vLLM, HuggingFace Transformers — dans des enclaves SGX sans recompilation, via un PAL (Platform Adaptation Layer) transparent. Pour les CVMs TDX et SEV-SNP, le projet Confidential Containers (CoCo) fournit l'intégration Kubernetes via Kata Containers, avec support natif AKS et GKE. NVIDIA Confidential Computing SDK expose les API pour exécuter des kernels CUDA dans des GPU confidentiels H100 en maintenant la confidentialité des poids lors du transfert CPU-GPU. Des frameworks d'inférence comme vLLM ont intégré des modes confidentiels avec RA-TLS en 2025-2026, simplifiant le déploiement de LLM open-source (LLaMA 3, Mistral, Phi-4) dans des environnements de confiance vérifiable sans code spécifique aux TEE côté applicatif.

Conclusion

Le **confidential computing LLM avec enclaves TEE** est passé en 2026 du stade de recherche académique à celui de déploiement industriel mature. Intel TDX et AMD SEV-SNP offrent des garanties cryptographiques robustes pour protéger les modèles propriétaires et les données utilisateurs contre des adversaires disposant d'accès privilégiés à l'infrastructure — administrateurs, opérateurs cloud, ou attaquants ayant compromis l'hyperviseur. L'attestation à distance, ancrée dans le silicium, élimine la nécessité de faire confiance aveuglément à un opérateur cloud et transforme la sécurité d'une promesse contractuelle en une garantie mathématiquement vérifiable.

Les challenges résiduels — overhead de performance de 8-20 %, attaques side-channel sur infrastructures multi-tenant, absence de protection intrinsèque contre les backdoors de modèles — sont des problèmes actifs pour lesquels les solutions progressent rapidement en 2026. Une stratégie de sécurité IA mature intègre le confidential computing comme couche défensive fondamentale, combinée à la vérification de la provenance des modèles, au monitoring

comportemental des sorties LLM et à la sécurisation des bases de données vectorielles. Les organisations qui adoptent cette approche aujourd'hui bâtissent une infrastructure IA auditable, réglementairement conforme NIS 2 et RGPD, et résiliente aux menaces internes comme aux fournisseurs potentiellement hostiles.