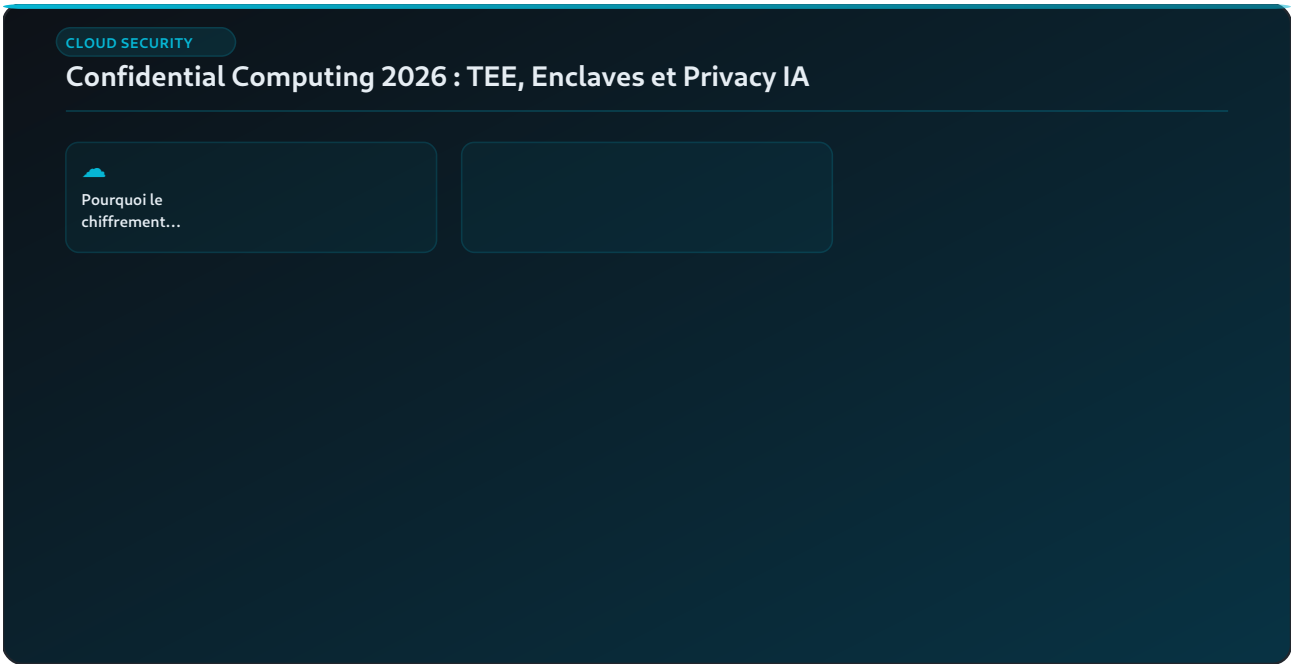
Confidential Computing 2026 : TEE, Enclaves et Privacy IA

> Guide confidential computing en 2026 — Intel SGX/TDX, AMD SEV-SNP, Arm CCA, enclaves pour LLM confidentiels, attestation et cas usage santé/finance

Le confidential computing répond à une question fondamentale que le chiffrement classique ne peut pas résoudre : comment protéger des données sensibles pendant leur traitement, pas seulement au repos ou en transit ? Les *Trusted Execution Environments (TEE)* — enclaves sécurisées dans le processeur lui-même — permettent d'exécuter du code et de traiter des données dans une zone isolée que ni le système d'exploitation, ni l'hyperviseur, ni même l'opérateur cloud ne peuvent lire ou modifier. En 2026, cette technologie sort des laboratoires de recherche pour devenir une brique d'infrastructure critique dans les déploiements d'IA en santé, finance et défense. Intel SGX, Intel TDX, AMD SEV-SNP, Arm CCA — les implémentations matérielles prolifèrent, les services cloud managés se démocratisent, et les cas d'usage IA en confidential computing deviennent réalité industrielle. Cet article décrypte les architectures TEE en 2026, leurs applications aux LLM confidentiels, et la roadmap pratique pour les organisations souhaitant sécuriser leurs charges de travail IA les plus sensibles.



Pourquoi le chiffrement classique ne suffit pas pour l'IA

Le chiffrement TLS protège les données en transit. Le chiffrement AES protège les données au repos sur le disque. Mais pendant le traitement — quand un modèle IA analyse des données médicales, quand un LLM traite des contrats confidentiels, quand un algorithme de scoring analyse des données financières — les données sont en clair dans la RAM et accessibles à tout acteur ayant les privilèges système suffisants.

Dans un cloud mutualisé, cela signifie que l'opérateur cloud (AWS, Azure, GCP) a techniquement accès aux données traitées par ses clients. Pour les données particulièrement sensibles — génomes, dossiers médicaux, secrets commerciaux, données de défense — cette exposition constitue un risque réglementaire et contractuel inacceptable. Le **confidential computing** adresse précisément ce vide en chiffrant les données en mémoire pendant leur traitement.

Architecture des Trusted Execution Environments (TEE)

Un *Trusted Execution Environment* est une zone d'exécution sécurisée isolée du reste du système, implémentée au niveau du processeur. Les données et le code qui y s'exécutent sont chiffrés par des clés gérées exclusivement dans le matériel — inaccessibles au système d'exploitation, à l'hyperviseur, et même aux administrateurs.

⚡ **Retour terrain**

La configuration par défaut des providers cloud est rarement sécurisée. J'applique systématiquement un benchmark CIS au premier audit : AWS CIS Benchmark, Azure Security Benchmark, ou GCP CIS Benchmark selon le cas. Sur les 50 derniers environnements cloud audités, aucun n'atteignait le score minimal de conformité CIS niveau 1 sans intervention préalable. Les findings les plus fréquents : logging CloudTrail/

Activity Log désactivé sur certaines régions, MFA non forcé sur les comptes root/admin, et SGs/NSGs avec des règles 0.0.0.0/0 en entrée.

Technologie	Fournisseur	Granularité	Overhead perf	Cas d'usage IA
Intel SGX	Intel	Application (enclave)	10-30%	Inférence LLM, scoring
Intel TDX	Intel	VM entière	3-10%	Entraînement, pipelines ML
AMD SEV-SNP	AMD	VM entière	3-8%	Pipelines ML, LLM serving
Arm CCA	Arm	Realm (domaine isolé)	1-5%	Edge IA, IoT industriel
Nvidia H100 CC	Nvidia	GPU + attestation	2-5%	LLM confidentiels GPU

Intel SGX : les enclaves applicatives

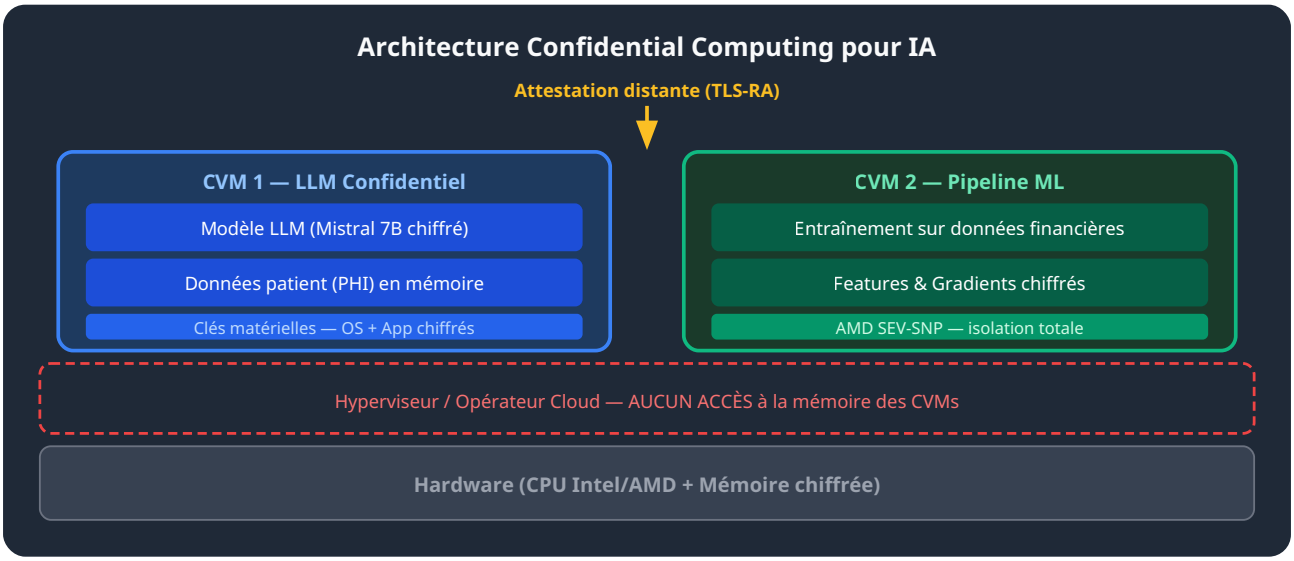
Intel Software Guard Extensions (SGX) est la technologie TEE la plus mature, disponible sur les processeurs Intel depuis 2015. SGX permet d'isoler des portions d'une application dans une *enclave SGX* — une région mémoire chiffrée par des clés matérielles. Le code et les données dans l'enclave sont inaccessibles au reste du système, y compris au kernel, aux hyperviseurs, et aux outils de débogage hardware.

La principale limitation de SGX est la taille mémoire des enclaves : historiquement limitée à 128 Mo (EPC - Enclave Page Cache), elle a été étendue dans SGX v2 mais reste contrainte. Pour l'inférence de petits modèles ou de modèles quantizés, SGX est viable. Pour des LLMs de plusieurs dizaines de milliards de paramètres, la contrainte mémoire est rédhibitoire sans techniques avancées de swap d'enclave — qui dégradent significativement les performances.

Intel TDX et AMD SEV-SNP : les Confidential VMs

Intel Trust Domain Extensions (TDX) et **AMD Secure Encrypted Virtualization-Secure Nested Paging (SEV-SNP)** adoptent une approche différente de SGX : plutôt d'isoler une application dans une enclave, ils isolent une *Virtual Machine (VM) entière*. La mémoire de la VM est chiffrée par des clés matérielles, invisible à l'hyperviseur et à l'opérateur cloud. Ces VM confidentielles (**Confidential VMs** ou CVM) permettent de faire tourner des charges de travail complètes — incluant un OS complet et tous ses processus — dans un environnement confidential.

L'avantage majeur par rapport à SGX : les Confidential VMs permettent de déployer des workloads IA existants sans modification du code. Un pipeline d'inférence LLM qui tourne sur une VM classique peut être migré sur une CVM AMD SEV-SNP avec peu ou pas de changements applicatifs. L'overhead de performance est limité à 3-10% selon les workloads — acceptable pour la plupart des cas d'usage IA en production.



Arm CCA : la révolution du confidential computing mobile et edge

Arm Confidential Compute Architecture (CCA), introduite dans l'architecture Armv9, étend le modèle TEE vers les appareils mobiles, les serveurs et les équipements edge. CCA introduit le

concept de *Realm* — un domaine d'exécution isolé analogue aux CVMs d'Intel TDX — géré par le *Realm Management Monitor (RMM)*, un composant firmware de confiance.

Pour les cas d'usage IA edge (inférence de modèles sur des équipements industriels, médicaux ou de défense), Arm CCA ouvre des perspectives inédites. Un modèle de détection de défauts déployé sur une chaîne de production peut s'exécuter dans un Realm, garantissant que le modèle propriétaire ne peut être extrait par un accès physique à l'équipement. L'overhead de performance sur Arm CCA est particulièrement faible (1-5%), ce qui le rend attractif pour des applications temps-réel sur le edge.

Nvidia H100 Confidential Computing : la GPU TEE

En 2023, Nvidia a introduit le support du **Confidential Computing sur GPU H100**, une innovation majeure pour l'IA. Les GPU H100 CC permettent d'exécuter des inférences LLM dans un environnement confidential où ni l'hôte, ni l'hyperviseur, ni l'opérateur cloud ne peuvent accéder aux données traitées sur le GPU. Le *Confidential CUDA* chiffre les transferts mémoire CPU-GPU et protège les données dans la VRAM du GPU.

L'intégration avec les CVMs Intel TDX ou AMD SEV-SNP crée une chaîne de confiance complète : données chiffrées en mémoire CPU (TDX/SEV-SNP), transferts CPU-GPU chiffrés, données chiffrées en VRAM H100. Cette architecture permet pour la première fois de déployer des LLMs de grande taille (plusieurs dizaines de milliards de paramètres) en mode confidential en production cloud. Des offres managées commencent à apparaître chez les hyperscalers en 2026 : Azure Confidential AI (H100 CC + AMD SEV-SNP), Google Cloud Confidential VM avec accélérateurs.

Attestation distante : le mécanisme de confiance fondamental

L'*attestation distante* (Remote Attestation) est le mécanisme qui permet à une application cliente de vérifier cryptographiquement qu'elle communique avec un TEE authentique exécutant le code

attendu. Sans attestation, le confidential computing serait inutile — rien ne garantirait que le code exécuté dans l'enclave est bien le code de confiance déclaré.

Le processus d'attestation : le TEE génère un **rapport d'attestation** signé par les clés matérielles du processeur, décrivant la configuration du TEE et le hash du code exécuté. Ce rapport est transmis au client via un service d'attestation tiers (Intel DCAP pour SGX, AMD Key Distribution Service pour SEV-SNP). Le client vérifie la signature du rapport contre les certificats racine d'Intel ou AMD, puis vérifie que le hash du code correspond au code de confiance attendu. Cette vérification, établie une fois en début de session, permet d'établir un canal TLS sécurisé directement avec le TEE — cryptographiquement prouvé comme authentique.

LLM confidentiels : cas d'usage santé et finance

Les deux secteurs qui bénéficient le plus immédiatement du confidential computing pour l'IA sont la santé et la finance, où les données traitées par les LLMs sont parmi les plus sensibles qui soient.

En **santé**, des hôpitaux et laboratoires pharmaceutiques déploient des LLMs pour analyser des dossiers médicaux électroniques, des résultats d'imagerie, ou des données génomiques. Ces données sont protégées par le RGPD, le Health Insurance Portability and Accountability Act (HIPAA) aux États-Unis, et les référentiels de certification des hébergeurs de données de santé (HDS en France). Un LLM médical exécuté dans une CVM AMD SEV-SNP sur un hébergeur certifié HDS permet de traiter des données patients avec une garantie cryptographique que l'opérateur cloud n'y a pas accès — bien au-delà des engagements contractuels traditionnels.

En **finance**, les cas d'usage confidential AI concernent principalement l'analyse de données de trading propriétaires, les modèles de risque incluant des données clients ultra-sensibles, et la détection de fraude sur des transactions confidentielles. Des banques françaises et européennes explorent activement ces architectures pour leurs projets d'IA générative sur données internes, avec un œil sur les exigences DORA (Digital Operational Resilience Act) qui imposent de nouvelles obligations de résilience et de traçabilité.

Federated Learning et confidential computing : une alliance puissante

Le *Federated Learning* — entraîner un modèle ML sur des données distribuées sans les centraliser — est souvent présenté comme une solution de confidentialité. Mais le federated learning classique expose toujours les gradients des modèles locaux, dont il est possible d'inférer des informations sur les données d'entraînement (gradient inversion attacks).

La combinaison **Federated Learning + Confidential Computing** résout ce problème : chaque client entraîne son modèle local dans un TEE, qui garantit que les gradients sont produits par le code d'entraînement déclaré sur des données réelles. Les gradients chiffrés sont agrégés dans un autre TEE (serveur d'agrégation confidentiel), invisible à l'orchestrateur central. Cette architecture, popularisée sous le nom de **Federated Learning with TEE (FL-TEE)**, est en cours de déploiement dans des consortiums de santé européens pour entraîner des modèles d'IA médicale collaborative sans jamais centraliser les données patients.

SMPC et homomorphic encryption : alternatives et compléments

Le confidential computing via TEE n'est pas la seule approche pour protéger les données en traitement. Deux autres techniques cryptographiques offrent des garanties différentes et peuvent se combiner avec les TEE.

Le *Secure Multi-Party Computation (SMPC)* permet à plusieurs parties de calculer conjointement une fonction sur leurs données respectives sans que personne ne voie les données des autres. Avec SMPC, un hôpital et une compagnie pharmaceutique peuvent entraîner collaborativement un modèle IA sans que ni l'un ni l'autre ne voit les données de l'autre — et sans TEE. L'*Homomorphic Encryption (HE)* va plus loin encore : elle permet d'effectuer des calculs directement sur des données chiffrées, dont le résultat, une fois déchiffré, est identique au calcul sur les données en clair. Pour l'inférence IA, HE permettrait d'envoyer des requêtes chiffrées à un modèle hébergé dans le cloud et de recevoir des résultats chiffrés — sans jamais déchiffrer côté serveur. En 2026, HE reste trop lente pour des LLMs (overhead de 1000x à

10000x), mais les bibliothèques comme SEAL (Microsoft), Concrete ML (Zama) et OpenFHE progressent rapidement.

Questions fréquentes sur le confidential computing

Quelle différence entre une VM chiffrée et une Confidential VM ?

Une VM chiffrée au repos (chiffrement du disque) protège les données quand la VM est éteinte, mais les données sont déchiffrées en mémoire lors de l'exécution — accessibles à l'hyperviseur et à l'opérateur cloud. Une **Confidential VM (CVM)** protège les données en mémoire pendant l'exécution via des clés matérielles gérées par le CPU. Même si l'hyperviseur dump la mémoire RAM de la CVM, il ne voit que des données chiffrées inutilisables. La CVM offre donc une protection contre un acteur de menace qui aurait compromis l'hyperviseur ou qui serait l'opérateur cloud lui-même.

Le confidential computing protège-t-il contre les side-channel attacks ?

Partiellement. Les TEE réduisent significativement la surface d'attaque, mais des vulnérabilités de type side-channel (Spectre, Meltdown, et leurs variantes pour SGX comme Foreshadow/L1TF) ont démontré qu'il est parfois possible d'inférer des informations depuis le contenu d'une enclave en observant des comportements du processeur (cache timing, power consumption). Intel et AMD ont continuellement corrigé ces vulnérabilités, mais le principe général reste : les TEE ne sont pas une protection parfaite contre un attaquant ayant accès physique au serveur et des ressources pour conduire des attaques side-channel sophistiquées. Pour la grande majorité des menaces réalistes (opérateurs cloud, attaquants réseau, administrateurs malveillants), les TEE offrent des garanties très solides.

Quel cloud provider offre le meilleur support du confidential computing pour l'IA ?

En 2026, **Microsoft Azure** offre l'écosystème le plus complet pour le confidential computing IA : Confidential VMs AMD SEV-SNP et Intel TDX, intégration avec Azure OpenAI Service en mode confidentiel, et Azure Confidential Computing SDK mature. **Google Cloud** propose des Confidential VMs AMD SEV-SNP et N2D bien intégrées avec Vertex AI. **AWS** propose AWS Nitro Enclaves (basé sur un hyperviseur propriétaire) et supporte les Confidential VMs AMD SEV-SNP sur certaines instances. Pour les déploiements en France avec des exigences de souveraineté, OVHcloud propose des Confidential VMs AMD SEV-SNP avec hébergement garanti sur le territoire français.

Comment intégrer l'attestation distante dans une application existante ?

L'intégration de l'attestation distante dans une application nécessite : côté serveur (dans la CVM), l'utilisation d'un SDK d'attestation spécifique à la technologie TEE (Intel DCAP pour TDX, AMD Attestation SDK pour SEV-SNP) pour générer les rapports d'attestation. Côté client, une bibliothèque de vérification d'attestation valide le rapport contre les certificats racine du fabricant. Des frameworks comme **Microsoft Azure Attestation** ou **Gramine Shielded Containers** abstraient une partie de cette complexité. Pour un projet démarrant, la solution la plus simple est d'utiliser une offre cloud managée (Azure Confidential Computing, GCP Confidential GKE) qui gère l'attestation nativement.

Quel est l'impact du confidential computing sur les performances des LLMs ?

L'impact dépend de la technologie TEE utilisée. Pour **Intel SGX** avec une enclave applicative, l'overhead peut atteindre 10-30% sur des opérations I/O intensives, mais les opérations de calcul pur (inférence IA) ont un overhead plus faible car les données en mémoire d'enclave sont déjà en clair une fois chargées. Pour les **Confidential VMs (TDX/SEV-SNP)**, l'overhead est de 3-10% pour des workloads IA typiques — pratiquement imperceptible par les utilisateurs. L'overhead GPU avec Nvidia H100 CC est estimé à 2-5% selon les workloads. Pour la plupart des déploiements IA en production, ces overheads sont acceptables et justifiés par le gain en sécurité.

À RETENIR

Points clés à retenir

TEE = protection en mémoire — le confidential computing protège les données pendant leur traitement, comblant le vide laissé par le chiffrement au repos et en transit.

Confidential VMs (TDX/SEV-SNP) pour l'IA — overhead faible (3-10%), déploiement sans modification du code, adapté aux LLMs et pipelines ML.

Nvidia H100 CC ouvre le GPU confidentiel — pour la première fois, des LLMs de grande taille peuvent s'exécuter en mode confidential sur GPU.

L'attestation distante est le mécanisme fondamental — sans attestation, impossible de prouver cryptographiquement qu'un TEE authentique exécute le bon code.

FL + TEE = formation collaborative confidentielle — entraîner des modèles sur des données distribuées sans les centraliser NI exposer les gradients.

OVHcloud pour la souveraineté française — Confidential VMs AMD SEV-SNP avec hébergement garanti en France pour les secteurs régulés.

La mise en œuvre du confidential computing pour vos charges IA nécessite une expertise combinant architecture cloud, sécurité hardware et développement ML. Notre équipe spécialisée en architecture IA sécurisée peut vous accompagner dans l'évaluation, le choix technologique et le déploiement. Pour les secteurs santé et finance, notre offre de conformité réglementaire IA intègre le confidential computing dans la stratégie de mise en conformité AI Act et RGPD. Consultez également notre expertise en sécurité des systèmes IA pour évaluer la robustesse de vos enclaves et CVMs. Pour les équipes souhaitant monter en compétences, notre programme de formation IA sécurisée couvre les fondamentaux du confidential computing appliqués à l'IA.

Ressources officielles : la Confidential Computing Consortium, consortium Linux Foundation regroupant Intel, AMD, Arm, Nvidia, Microsoft, Google et Red Hat pour standardiser le confidential computing, et la documentation technique Intel TDX.

Déployer un LLM confidentiel avec Gramine : guide pratique

Gramine (anciennement Graphene) est un framework open-source qui permet d'exécuter des applications Linux non modifiées dans une enclave Intel SGX ou sur des Confidential VMs. Pour déployer un LLM confidentiel avec Gramine, le workflow typique comprend : configurer le manifest Gramine qui décrit les fichiers, bibliothèques et ressources réseau accessibles depuis l'enclave, construire l'image de l'application Gramine (gramine-sgx-sign), déployer sur une instance cloud SGX-compatible, et vérifier l'attestation.

La grande valeur de Gramine pour les équipes IA est qu'il permet de sécuriser des frameworks existants (Python, PyTorch, TensorFlow, HuggingFace Transformers) sans modifier le code. Un

script Python d'inférence LLM tourne identiquement dans Gramine SGX avec des garanties confidential. Des projets comme **Gramine CI/CD** intègrent la construction et la vérification d'attestation dans les pipelines de déploiement, automatisant la chaîne de confiance de bout en bout.

Open Enclave SDK et Enarx : standards de portabilité

L'un des défis du confidential computing est la fragmentation des APIs entre SGX, TDX, SEV-SNP et Arm CCA. Des projets de standardisation émergent pour permettre aux développeurs d'écrire du code TEE portable entre plateformes. **Open Enclave SDK** (Microsoft, open-source) offre une couche d'abstraction unifiée supportant SGX et TDX. **Enarx** (Red Hat, Confidential Computing Consortium) va plus loin avec une abstraction complète des backends TEE, permettant de déployer le même WebAssembly bytecode sur SGX, SEV-SNP ou TDX sans modification.

Pour les équipes IA souhaitant construire des applications confidential portables, *Enarx + WebAssembly* est une combinaison prometteuse : le modèle d'inférence et le serveur d'inférence sont compilés en WebAssembly, puis exécutés dans un Enarx Keep (une instance TEE abstraite) sur le backend matériel disponible. La performance WebAssembly pour l'inférence IA s'améliore rapidement avec des runtimes spécialisés comme WasmEdge qui supporte ONNX et les accélérateurs matériels.

Anecdote terrain : confidential computing pour la génomique

En 2025, un consortium de cinq hôpitaux universitaires français a déployé un système de confidential AI pour l'analyse génomique collaborative. L'objectif : entraîner un modèle de prédiction du risque oncologique sur les données génomiques de 50 000 patients des cinq

hôpitaux, sans qu'aucun hôpital ne transmette ses données à un serveur central — une exigence légale et éthique non négociable dans ce contexte.

L'architecture déployée combinait Federated Learning avec des enclaves TEE (AMD SEV-SNP sur OVHcloud HDS) pour chaque site participant. Chaque hôpital entraînait son modèle local dans une CVM SEV-SNP, garantissant que les gradients partagés ne pouvaient pas être interceptés par l'opérateur cloud. L'agrégation des gradients se faisait dans une CVM dédiée sur un serveur neutre, dont l'attestation était vérifiable par tous les participants. Résultat : un modèle avec une performance de prédiction supérieure de 23% à un modèle entraîné sur les données d'un seul hôpital, obtenu sans jamais centraliser les données patients. Ce projet illustre comment le confidential computing transforme des collaborations scientifiques auparavant impossibles en réalité opérationnelle.

Politique de sécurité pour les enclaves : définir la Trusted Computing Base

La *Trusted Computing Base (TCB)* d'une enclave TEE définit l'ensemble des composants dont la compromission entraînerait une violation des garanties de sécurité. Minimiser la TCB est un principe fondamental de la sécurité TEE : chaque ligne de code dans la TCB est une ligne de code susceptible de contenir des vulnérabilités.

Pour une enclave SGX, la TCB inclut : le code de l'application dans l'enclave, les bibliothèques SDK SGX, et le microcode du processeur Intel. Pour une CVM TDX/SEV-SNP, la TCB inclut le kernel de la VM, le code de l'application, et le firmware du processeur. Les bonnes pratiques de design d'enclave : minimiser le code dans la TCB en déplaçant le maximum de logique hors de l'enclave (dans la partie "non confidentielle" de l'application), auditer rigoureusement le code dans la TCB, utiliser des langages mémoire-safe (Rust est largement préféré à C pour le code TEE), et documenter explicitement les hypothèses de sécurité.

Confidential computing et cloud souverain français

La France dispose d'un cadre spécifique pour la souveraineté des données cloud, formalisé par la qualification **SecNumCloud** de l'ANSSI. En 2026, la version 3.2 de SecNumCloud intègre des exigences sur l'immunité aux lois extraterritoriales — un point crucial pour les données les plus sensibles de l'État et des opérateurs d'importance vitale (OIV).

Le confidential computing joue un rôle important dans cette stratégie de souveraineté. OVHcloud, qualifié SecNumCloud, propose des Confidential VMs AMD SEV-SNP permettant aux clients de déployer des charges IA avec des garanties cryptographiques que l'opérateur (y compris OVHcloud lui-même) n'a pas accès aux données en traitement. Cette garantie technique complète les engagements contractuels et organisationnels. Pour les OIV et les administrations françaises traitant des données IA sensibles (analyse de renseignement, modèles de prévention de fraude fiscale, IA médicale de l'AP-HP), le couple SecNumCloud + Confidential VMs représente le niveau de protection le plus élevé actuellement disponible en cloud public.

Opinion : le confidential computing va devenir le standard par défaut pour l'IA en secteurs régulés

Voici une position tranchée : d'ici 2028, le déploiement de LLMs sur des données sensibles sans confidential computing sera considéré comme une non-conformité de sécurité dans les secteurs de la santé, de la finance et de la défense — au même titre que le déploiement d'une application web sans HTTPS l'est aujourd'hui. Ce n'est pas une spéculation : la trajectoire réglementaire, l'évolution des menaces, et la maturité technique convergent dans cette direction.

L'AI Act (Art. 15) exige que les systèmes IA à haut risque atteignent un niveau approprié de cybersécurité. DORA impose des mesures de protection contre les accès non autorisés dans les systèmes critiques financiers. Les référentiels HDS et SecNumCloud évoluent vers des exigences de protection en mémoire. Et technologiquement, le support natif du confidential computing dans les principales plateformes cloud et les processeurs de nouvelle génération réduit le coût et la complexité de déploiement à un niveau acceptable pour des projets de taille moyenne. Les

équipes IA qui investissent aujourd'hui dans la maîtrise du confidential computing se positionnent avantageusement pour la prochaine vague de déploiements IA régulés.

Coûts et dimensionnement d'une infrastructure confidential AI

Construire une infrastructure confidential computing pour l'IA nécessite de prendre en compte plusieurs postes de coût. Côté compute, les instances cloud avec Confidential VMs sont facturées en prime par rapport aux VMs classiques : compter environ 10-20% de surcoût pour des instances AMD SEV-SNP, et une prime plus élevée pour les configurations avec Nvidia H100 CC (actuellement disponibles sur un nombre limité de régions).

Côté développement, l'adaptation d'une application IA existante au confidential computing nécessite typiquement 2-4 semaines pour un déploiement CVM (minimal : choisir le type d'instance, configurer l'attestation, adapter les scripts de déploiement) à 2-4 mois pour un déploiement SGX avec enclave applicative (refactoring de l'architecture, gestion de la TCB, tests de performance). La maintenance supplémentaire est limitée : les mises à jour de microcode et de firmware sont gérées par le fournisseur cloud, les updates du code dans la CVM suivent le processus habituel avec une étape de re-attestation.

Roadmap confidential computing pour une organisation

Pour une organisation souhaitant adopter le confidential computing pour ses charges IA, une roadmap en trois phases permet une montée en maturité progressive. La **Phase 1 - Expérimentation** (1-3 mois) : identifier les 2-3 cas d'usage IA avec les données les plus sensibles, déployer une preuve de concept sur une CVM AMD SEV-SNP dans le cloud existant, mesurer l'overhead de performance réel sur des workloads représentatifs, former 2-3 ingénieurs aux concepts TEE et à l'attestation.

La **Phase 2 - Déploiement pilote** (3-6 mois) : déployer le premier cas d'usage en production sur une Confidential VM, intégrer la vérification d'attestation dans le client applicatif, auditer la TCB avec un expert sécurité externe, documenter les procédures opérationnelles. La **Phase 3 - Généralisation** (6-18 mois) : étendre le confidential computing à l'ensemble des charges IA à données sensibles, explorer l'adoption de Federated Learning + TEE pour les projets collaboratifs, contribuer aux standards internes de sécurité IA pour systématiser l'usage du confidential computing dans les nouveaux projets. Notre équipe spécialisée en architecture cloud sécurisée peut accompagner chaque phase de cette roadmap.

Intégration avec les outils MLOps existants

L'une des questions pratiques les plus fréquentes sur le confidential computing pour l'IA est son intégration avec les pipelines MLOps existants. La bonne nouvelle : la plupart des outils MLOps standards (MLflow, Kubeflow, DVC, Weights&Biases) fonctionnent sans modification dans des Confidential VMs AMD SEV-SNP ou Intel TDX, car ces technologies protègent l'environnement d'exécution au niveau de la VM sans imposer de contraintes applicatives.

Pour **Kubernetes**, des projets comme **Constellation** (Edgeless Systems) permettent de déployer des clusters Kubernetes entièrement confidentiels, où chaque nœud est une CVM attestée. Les pods Kubernetes s'exécutent dans des environnements confidentiels sans modification des manifests Kubernetes existants. Cette compatibilité avec l'écosystème cloud-native standard est un facteur d'adoption majeur : les équipes IA n'ont pas à réécrire leur infrastructure MLOps pour bénéficier du confidential computing. Notre équipe SOC surveille activement les flux d'attestation des enclaves dans les environnements de production pour détecter toute anomalie de configuration.

Confidential AI et réglementation : AI Act, DORA, SecNumCloud

Le confidential computing se positionne comme une réponse technique aux exigences croissantes des réglementations IA et données. L'**AI Act (Art. 15)** exige que les systèmes IA à haut risque atteignent un niveau approprié de cybersécurité, notamment une résistance aux tentatives d'accès non autorisé. Les Confidential VMs répondent directement à cette exigence en garantissant cryptographiquement que même un opérateur cloud compromis ne peut accéder aux données de traitement.

DORA (Digital Operational Resilience Act), applicable depuis janvier 2025, impose aux entités financières européennes des exigences renforcées sur la sécurité des systèmes d'information tiers (notamment cloud). Le confidential computing répond à l'Article 30 de DORA qui exige des mesures contractuelles et techniques pour garantir que les tiers prestataires n'accèdent aux données que dans la mesure strictement nécessaire — une garantie que seul le chiffrement en mémoire peut fournir de façon cryptographiquement prouvable, au-delà des simples engagements contractuels.

Mesures de performance : benchmarks confidential computing pour LLMs

Des benchmarks publiés par Microsoft Research, OVHcloud et des équipes académiques en 2025 quantifient l'overhead du confidential computing pour les charges IA typiques. Sur des benchmarks d'inférence LLM (Llama 3 8B, Mistral 7B) sur des CVMs AMD SEV-SNP, l'overhead de latence est de **4-7%** par rapport à des VMs classiques de même spécification, avec un overhead mémoire négligeable (la mémoire chiffrée est gérée transparentement par le CPU).

Pour l'entraînement de modèles sur des CVMs Intel TDX, l'overhead est similaire : 5-9% sur des benchmarks MLPerf. Les opérations les plus impactées sont les I/O intensives (lecture de datasets depuis le stockage) où l'overhead peut atteindre 15-20% sur des configurations avec chiffrement I/O activé. La solution standard : stocker les données d'entraînement dans un système de fichiers en mémoire (tmpfs) pré-chargé au démarrage de la CVM depuis un stockage chiffré externe, évitant ainsi les I/O répétés pendant l'entraînement. Pour l'inférence GPU avec

Nvidia H100 CC, l'overhead mesuré est de 2-4% sur des benchmarks de throughput de tokens — imperceptible en production.

Gestion des secrets dans les enclaves : HSM et Key Management

Un aspect critique du confidential computing est la gestion des clés cryptographiques qui protègent les données dans les enclaves. Contrairement à un déploiement cloud classique où les clés peuvent être stockées dans un Key Management Service (AWS KMS, Azure Key Vault), les enclaves TEE nécessitent une approche différente pour préserver les garanties de confidentialité.

Le pattern recommandé est le *Sealed Storage* (Intel SGX) ou l'équivalent pour les CVMs : les secrets sont chiffrés par des clés dérivées du CPU lui-même (MRSIGNER, MRENCLAVE pour SGX), de sorte qu'ils ne peuvent être déchiffrés que par la même enclave sur le même CPU. Pour les cas d'usage nécessitant une portabilité (migrer une enclave sur un autre serveur), des **Hardware Security Modules (HSM)** connectés à l'enclave via un canal attesté permettent de stocker et distribuer les clés de façon sécurisée. Des services comme **Azure Managed HSM** ou **AWS CloudHSM** supportent des protocoles d'intégration spécifiques aux enclaves.

Supply chain de confiance pour les modèles IA confidentiels

Le déploiement d'un LLM confidentiel soulève une question souvent négligée : comment s'assurer que le modèle lui-même n'a pas été compromis avant son chargement dans l'enclave ? Si un attaquant peut substituer un modèle malveillant à un modèle légitime lors du déploiement, les garanties du TEE sont ineffectives — l'enclave exécute fidèlement le mauvais modèle.

La réponse est une *supply chain de confiance pour les modèles* : chaque modèle doit être signé par son producteur, et l'intégrité de la signature doit être vérifiée avant chargement dans l'enclave. Des standards émergent pour les **Model Cards sécurisées** incluant des signatures cryptographiques (Sigstore, in-toto), permettant de tracer l'origine et l'intégrité d'un modèle

depuis son entraînement jusqu'à son déploiement. Pour les LLMs open-source hébergés sur HuggingFace, des projets pilotes de signature de modèles (Model Signing) permettent aux utilisateurs de vérifier que le modèle téléchargé correspond bien à celui publié par l'équipe de recherche. Cette approche "DevSecOps pour les modèles IA" sera vraisemblablement exigée par les futures versions des standards AI Act.

Confidential computing dans les architectures multi-cloud

Les grandes organisations opèrent souvent dans des environnements multi-cloud, ce qui complexifie la mise en œuvre du confidential computing. Les clés matérielles sont liées au processeur d'un fournisseur cloud spécifique — une enclave SGX sur AWS ne peut pas utiliser les clés d'une enclave SGX sur Azure. Deux approches permettent de gérer cette complexité.

La première est d'adopter un **Confidential Key Management Service (KMS) portable** : un service de gestion des clés lui-même s'exécutant dans une enclave, accessible depuis des CVMs sur différents clouds via des protocoles d'attestation standardisés. Des projets open-source comme **Confidant** (Lyft) ou **SPIFFE/SPIRE** permettent de gérer des identités et des secrets dans des environnements multi-cloud avec des garanties de confidentialité. La deuxième approche est de limiter le confidential computing à un cloud primaire "de confiance" et de n'y traiter que les données les plus sensibles, en déployant les charges moins critiques sur d'autres clouds sans contrainte TEE.

Formation et certification en confidential computing

Le marché de la formation en confidential computing est encore immature, mais des ressources sérieuses existent pour les équipes souhaitant développer ces compétences. Intel propose une **formation officielle Intel SGX SDK** (disponible en ligne), couvrant le développement d'enclaves et la gestion des attestations. AMD dispose de documentations techniques détaillées

sur SEV-SNP et d'un support développeur actif. Le **Confidential Computing Consortium** (CCC) publie des guides techniques et des use cases référence.

Pour les certifications professionnelles, le domaine est en cours de structuration. Des certifications comme CISSP, CISM ou AWS Security Specialist ne couvrent pas encore le confidential computing de façon approfondie. Des certifications spécifiques commencent à émerger de la part de fournisseurs cloud (Azure Confidential Computing Specialist). En France, des formations courtes proposées par des organismes comme l'ANSSI-académie ou des écoles spécialisées en cybersécurité intègrent progressivement le confidential computing dans leurs curricula. Notre équipe propose des ateliers pratiques confidential computing pour les équipes DevSecOps souhaitant maîtriser ces technologies dans un contexte de déploiement IA réel.

L'avenir du confidential computing : vers une intégration transparente

La trajectoire du confidential computing va vers une intégration de plus en plus transparente dans les workflows de développement et de déploiement. Plusieurs signaux confirment cette tendance. Du côté des processeurs, les fabricants intègrent les capacités TEE dans des générations de plus en plus larges de chips — les Confidential VMs deviennent disponibles sur davantage de types d'instances cloud, réduisant les compromis performance/coût.

Du côté des orchestrateurs, Kubernetes supporte nativement les Confidential VMs comme classe d'exécution via les node types confidentiels, et des projects comme **Kata Containers** permettent de faire tourner des pods Kubernetes dans des CVMs sans modification des manifests. Du côté des standards, l'émergence de l'**IETF RATS (Remote ATtestation procedureS)** standardise les protocoles d'attestation distante pour une interopérabilité entre fournisseurs. D'ici 2028, le confidential computing ne sera plus une technologie spécialisée nécessitant une expertise dédiée : ce sera un paramètre de déploiement dans les outils Infrastructure as Code (Terraform, Pulumi) — aussi simple à activer qu'un security group ou un chiffrement de volume EBS.

Cas d'usage défense et renseignement : IA dans des environnements très classifiés

Le domaine de la défense et du renseignement représente le segment où les exigences de confidentialité pour l'IA sont les plus strictes. Les forces armées et les services de renseignement déploient des systèmes d'IA pour l'analyse d'images satellite, la fusion de données de capteurs, l'aide à la décision tactique — des applications où la moindre fuite d'information peut avoir des conséquences stratégiques. En France, le Commandement du Numérique de Défense (COMNUM) et la DGA (Direction Générale de l'Armement) explorent le confidential computing pour permettre le déploiement d'IA sur des infrastructures cloud tout en maintenant des garanties de sécurité compatibles avec les systèmes d'information classifiés.

Les défis spécifiques de la défense incluent : la nécessité de travailler avec des données classifiées qui ne peuvent pas quitter des réseaux physiquement isolés (air-gapped), les exigences de certification spécifiques (CSEC, Crypto Aggréé) qui vont au-delà des standards commerciaux, et la nécessité d'audit de sécurité par des organismes habilités (ANSSI pour les systèmes sensibles). Le confidential computing hardware est un élément de réponse, mais il doit s'intégrer dans un cadre de sécurité globale incluant des HSM certifiés EAL4+, des protocoles d'attestation adaptés aux réseaux fermés, et des processus de qualification nationale.

Sécuriser les modèles de propriété intellectuelle avec les TEE

Au-delà de la protection des données clients, le confidential computing résout un problème de plus en plus critique pour les éditeurs de modèles IA : protéger leurs modèles propriétaires contre l'extraction par les clients ou les opérateurs cloud. Un LLM représente des investissements de plusieurs millions d'euros en données d'entraînement et en compute — un actif qu'un éditeur souhaite légitimement protéger.

Avec les TEE, un éditeur peut déployer son modèle dans une enclave où le modèle est chiffré par des clés matérielles inaccessibles au client-déployeur. Le client peut utiliser le modèle via une API (envoyer des inputs, recevoir des outputs), mais ne peut pas extraire les poids du modèle,

auditer son architecture interne, ou distiller un modèle concurrent. Cette propriété du **modèle confidentiel** ouvre un nouveau modèle commercial pour les éditeurs IA : déployer des modèles propriétaires chez les clients avec une protection matérielle de la propriété intellectuelle, sans nécessiter une relation de confiance organisationnelle avec le client. Des startups comme **Opaque Systems** ou **Fortanix** commercialisent exactement ce type de solution pour les entreprises qui veulent monétiser leurs modèles IA de façon sécurisée.

Audit de sécurité d'une enclave TEE : méthodologie

Un audit de sécurité d'une enclave TEE nécessite une méthodologie spécifique, différente d'un audit applicatif classique. Les vecteurs d'attaque à analyser incluent : les side-channel attacks (cache timing, power analysis), les attaques sur l'interface ocall/ecall (les points d'entrée/sortie de l'enclave), les vulnérabilités dans les bibliothèques importées dans l'enclave, la qualité de la génération de nombres aléatoires dans l'enclave (critique pour les opérations cryptographiques), et la robustesse du processus d'attestation contre des faux rapports d'attestation.

Notre équipe de sécurité offensive IA inclut des spécialistes de la sécurité hardware et des TEE, capables de conduire des audits approfondis des enclaves SGX/TDX/SEV-SNP incluant une analyse des side-channel risks, un fuzzing des interfaces ocall, et une revue du code dans la TCB. Un audit TEE complet représente typiquement 2-4 semaines pour une enclave applicative de taille moyenne, avec un livrable incluant un rapport de vulnérabilités classifiées CVSS et des recommandations de remédiation priorisées. Pour les organisations déployant des modèles IA dans des enclaves avec des données de santé ou des données financières réglementées, cet audit est une étape indispensable avant le passage en production.

Questions pratiques sur la mise en œuvre du confidential computing

Quelle est la différence entre confidential computing et zero-trust architecture ?

Ces deux approches sont complémentaires mais portent sur des couches différentes. Zero-trust est un modèle d'architecture réseau et d'identité qui suppose que personne n'est de confiance par défaut — chaque accès est authentifié et autorisé explicitement. Le confidential computing est une garantie hardware sur la confidentialité des données pendant leur traitement, indépendamment des politiques d'accès réseau. Un déploiement IA pleinement sécurisé combine les deux : zero-trust pour contrôler qui peut soumettre des requêtes à l'enclave, et confidential computing pour garantir que les données traitées dans l'enclave restent confidentielles même vis-à-vis des administrateurs infrastructure. L'attestation distante du TEE s'intègre naturellement dans un framework zero-trust comme preuve d'intégrité du workload.

Comment tester son application dans un TEE avant de déployer en production ?

Les outils de simulation TEE permettent de développer et tester des applications d'enclave sans avoir accès à un matériel SGX ou SEV-SNP. Intel SGX SDK inclut un mode simulation complet. Pour les CVMs TDX/SEV-SNP, QEMU supportent des modes d'émulation permettant de tester le comportement applicatif sans les garanties cryptographiques réelles. Pour les tests fonctionnels et de performance, des instances cloud de développement à bas coût (Azure DCsv3 pour SGX, instances AMD confidentielles sur AWS ou Azure) permettent de valider le comportement réel sur matériel. Un pipeline CI/CD complet pour une application TEE inclut typiquement : tests unitaires en simulation SGX, tests d'intégration sur instance cloud TEE de développement, vérification de l'attestation, benchmarks de performance, et déploiement progressif en production.