AYI NEDJIMI Consultants

Confidential Computing et IA : Entraîner et Inférer dans

Catégorie : Intelligence Artificielle Lecture : 12 min Publié le : 28/02/2026 Auteur : Ayi NEDJIMI

TEE (Intel TDX, AMD SEV-SNP, ARM CCA) pour l'IA : inférence confidentielle, entraînement multi-parties sécurisé,. Guide expert avec méthodologies et.

Table des Matières

Votre organisation est-elle prête à faire face aux attaques basées sur l'IA ? TEE (Intel TDX, AMD SEV-SNP, ARM CCA) pour l'IA : inférence confidentielle, entraînement multi-parties sécurisé,. Guide expert avec méthodologies et. Dans un contexte où l'intelligence artificielle transforme les pratiques de cybersécurité, la maîtrise de ia confidential computing enclaves securisees devient un avantage stratégique pour les équipes techniques. Nous abordons notamment : table des matières, 1 introduction au confidential computing pour l'ia et 2 technologies tee (intel tdx, amd

sev-snp, arm cca). Les professionnels y trouveront des recommandations actionnables, des commandes prêtes à l'emploi et des stratégies de mise en œuvre adaptées aux environnements d'entreprise.

1 Introduction au Confidential Computing pour l'IA

La protection des données est traditionnellement assurée selon trois états : **at rest** (chiffrement de stockage), **in transit** (TLS/mTLS) et **in use** (données en mémoire pendant le traitement). Si les deux premiers états bénéficient de solutions matures et largement déployées, la protection des données en cours de traitement reste le maillon faible. Le **Confidential Computing** résout ce problème en utilisant des environnements d'exécution de confiance matériels (Trusted Execution Environments, TEE) qui protègent les données en mémoire contre tout accès non autorisé, y compris de la part de l'opérateur de l'infrastructure (cloud provider, administrateur système).

L'intersection entre Confidential Computing et intelligence artificielle ouvre des possibilités considérables. Les organisations hésitent souvent à déployer des modèles IA dans le cloud pour des raisons de confidentialité : les données d'entraînement peuvent contenir des informations personnelles soumises au RGPD, les prompts utilisateurs révèlent des informations métier sensibles, et les poids du modèle constituent de la propriété intellectuelle de haute valeur. Le Confidential Computing permet de **déployer l'IA dans des enclaves sécurisées** où ni le cloud provider ni aucun administrateur ne peut accéder aux données en cours de traitement, aux prompts des utilisateurs, ni aux poids du modèle. Cette garantie est assurée par le matériel et vérifiable par attestation cryptographique.

Le **Confidential Computing Consortium** (CCC), fondé par la Linux Foundation en 2019 et regroupant Intel, AMD, ARM, Microsoft, Google, Meta et NVIDIA, pilote la standardisation des interfaces et des protocoles. En 2026, le marché du Confidential Computing pour l'IA connaît une croissance explosive, portée par les exigences réglementaires (RGPD, AI Act, HIPAA) et les cas d'usage en santé, finance et défense où la confidentialité des données est non négociable.

Définition clé : Le **Confidential Computing** protège les données en cours de traitement (in use) en utilisant des environnements d'exécution matériels isolés (TEE). Les données et le code à l'intérieur du TEE sont protégés contre tout accès externe — y compris du système d'exploitation, de l'hyperviseur et de l'opérateur de l'infrastructure — avec des garanties vérifiables par attestation cryptographique.

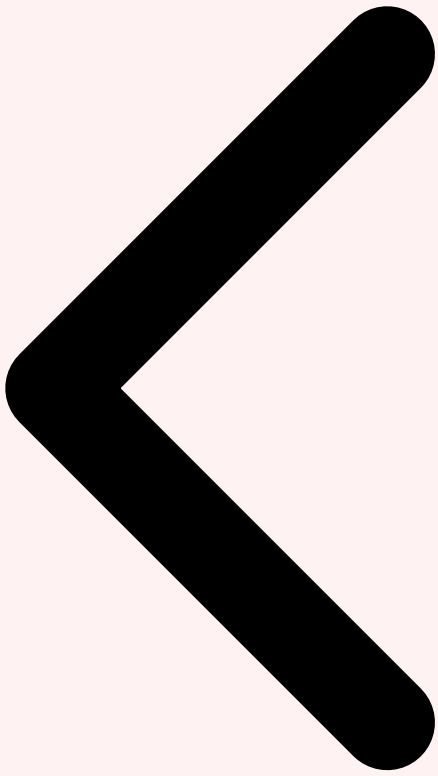
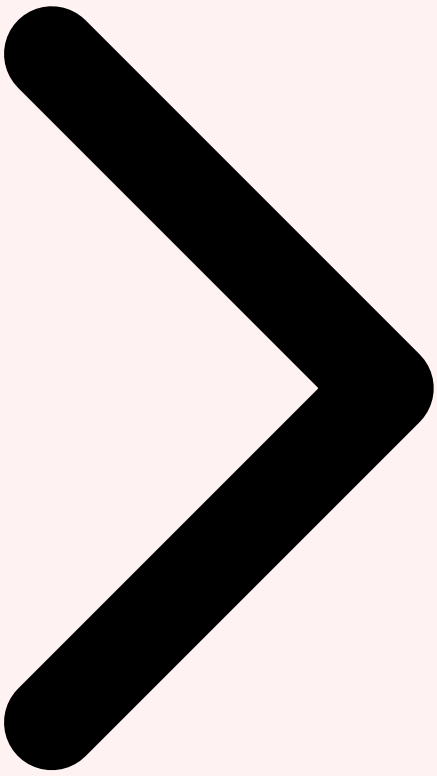


Table des Matières Introduction Technologies TEE



Critere	Description	Niveau de risque
Confidentialite	Protection des donnees d'entrainement et des prompts	Eleve
Integrite	Fiabilite des sorties et detection des hallucinations	Critique
Disponibilite	Resilience du service et gestion de la charge	Moyen
Conformite	Respect du RGPD, AI Act et politiques internes	Eleve

2 Technologies TEE (Intel TDX, AMD SEV-SNP, ARM CCA)

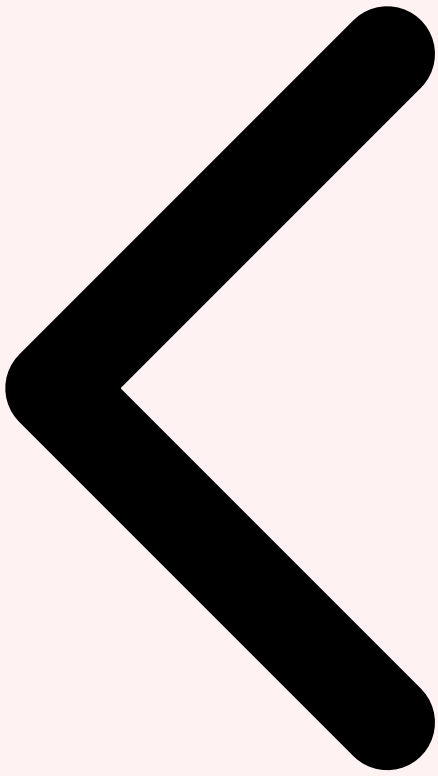
Trois technologies TEE majeures se partagent le marché en 2026, chacune avec des caractéristiques distinctes. **Intel Trust Domain Extensions (TDX)**, successeur d'Intel SGX, fournit une isolation au niveau de la machine virtuelle plutôt qu'au niveau de l'application. TDX crée des *Trust Domains* (TD) — des VMs complètes dont la mémoire est chiffrée par le processeur avec des clés matérielles inaccessibles à l'hyperviseur. L'avantage majeur de TDX est la compatibilité applicative : tout logiciel existant fonctionne dans un TD sans modification, éliminant le besoin de porter les applications dans un SDK spécialisé comme

c'était le cas avec SGX. TDX est disponible sur les processeurs Intel Xeon de 4ème génération (Sapphire Rapids) et suivants, avec une mémoire protégée pouvant atteindre plusieurs téraoctets.

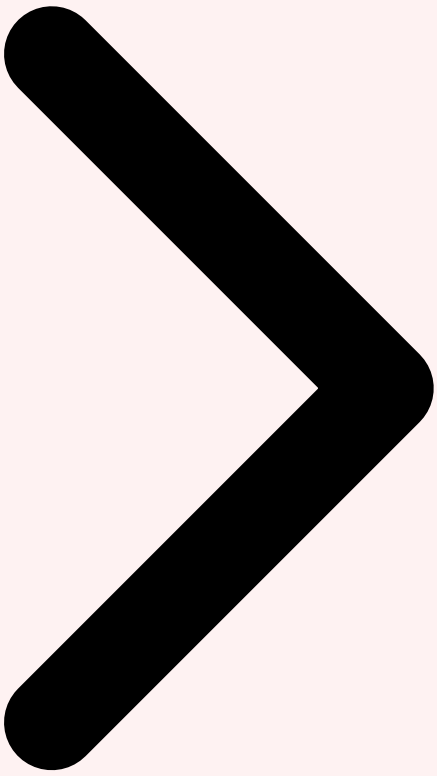
AMD Secure Encrypted Virtualization-Secure Nested Paging (SEV-SNP) est l'implémentation AMD de la VM confidentielle. SEV-SNP chiffre la mémoire de chaque VM avec des clés AES-256 gérées par un processeur de sécurité dédié (AMD Secure Processor, ASP). SNP ajoute l'intégrité mémoire (protection contre les attaques de remapping) et l'attestation cryptographique au SEV de base. AMD SEV-SNP est disponible sur les processeurs EPYC de 3ème génération (Milan) et suivants, et est largement déployé chez les cloud providers (Azure, AWS, GCP). L'un des avantages d'AMD SEV-SNP est sa capacité à protéger de très grands espaces mémoire (jusqu'à 509 clés de chiffrement simultanées), ce qui le rend particulièrement adapté aux workloads IA nécessitant plusieurs dizaines de gigaoctets de mémoire. Pour approfondir, consultez Sécurité LLM Adversarial : Attaques, Défenses et Bonnes.

ARM Confidential Compute Architecture (CCA), annoncé avec ARMv9, étend le modèle de sécurité ARM TrustZone avec des *Realms* — des environnements d'exécution isolés gérés par un *Realm Management Monitor* (RMM) matériel. CCA est particulièrement pertinent pour l'IA edge et mobile, où les modèles sont déployés sur des dispositifs ARM (smartphones, IoT, véhicules autonomes). **NVIDIA Confidential Computing**, via les GPU H100/H200 avec le mode *CC-On*, étend les garanties TEE au GPU. Les données et le code du modèle dans la mémoire GPU (HBM) sont chiffrés et protégés contre l'accès par l'hôte. Cette innovation est fondamentale pour l'IA confidentielle car les workloads ML sont massivement exécutés sur GPU.

- ▹**Intel TDX :** isolation au niveau VM, compatibilité applicative totale, mémoire chiffrée multi-To
- ▹**AMD SEV-SNP :** chiffrement AES-256, intégrité mémoire, attestation, large déploiement cloud
- ▹**ARM CCA :** Realms isolés, pertinent pour IA edge/mobile sur dispositifs ARMv9
- ▹**NVIDIA CC :** GPU H100/H200 en mode confidentiel, chiffrement HBM, essentiel pour ML



Introduction Technologies TEE Inférence confidentielle



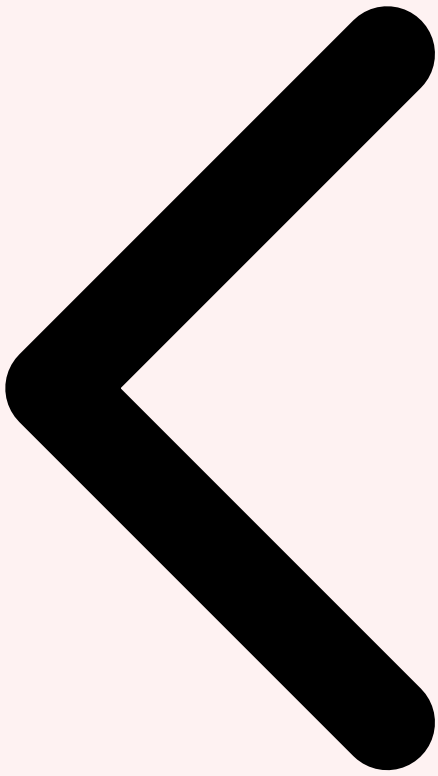
Notre avis d'expert
Chez Ayi NEDJIMI Consultants, nous constatons que la majorité des organisations sous-estiment les risques liés aux modèles de langage déployés en production. La sécurité des LLM ne se limite pas au prompt engineering : elle exige une approche systémique couvrant les embeddings, les pipelines de données et les mécanismes de contrôle d'accès aux API.

3 Inférence confidentielle

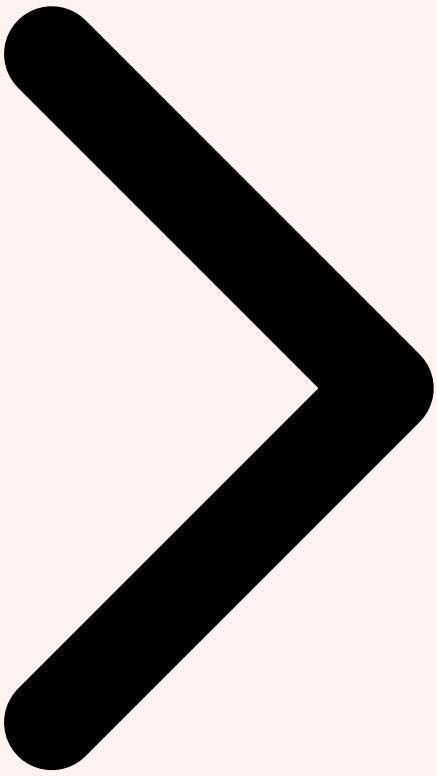
L'**inférence confidentielle** permet d'exécuter un modèle IA sur des données utilisateur sans que quiconque — ni l'opérateur du service, ni le cloud provider, ni un attaquant ayant compromis l'infrastructure — ne puisse accéder aux données d'entrée, aux résultats de l'inférence ou aux poids du modèle. Ce cas d'usage est fondamental pour les applications IA traitant des données hautement sensibles : diagnostic médical à partir d'imagerie, analyse de documents juridiques confidentiels, traitement de données financières, ou interrogation de bases de connaissances classifiées.

L'architecture d'inférence confidentielle typique déploie le modèle et le moteur d'inférence (vLLM, TGI, TensorRT-LLM) à l'intérieur d'un TEE (VM confidentielle TDX ou SEV-SNP). Les requêtes utilisateur arrivent via un canal TLS terminé à l'intérieur du TEE — le cloud provider ne voit que du trafic chiffré. Les GPU confidentiels NVIDIA (H100 CC-On) chiffrent les données en transit entre le CPU et le GPU via un lien PCIe sécurisé et chiffrent la mémoire HBM du GPU. Avant d'envoyer ses données, l'utilisateur peut vérifier l'attestation du TEE pour confirmer que le code attendu (modèle + moteur d'inférence + configuration) s'exécute bien dans un environnement confidentiel non modifié.

Apple a implémenté ce concept à grande échelle avec **Private Cloud Compute (PCC)**, annoncé en 2024 pour Apple Intelligence. PCC exécute les requêtes IA des utilisateurs Apple dans des enclaves sécurisées basées sur des puces Apple Silicon avec Secure Enclave, avec des garanties d'attestation publique et de non-rétention des données. **Azure Confidential AI** propose des VMs confidentielles (DCsv3, DCdsv3) avec GPU NVIDIA H100 en mode confidentiel pour le déploiement de modèles IA. **Google Cloud Confidential Space** offre un environnement similaire basé sur AMD SEV-SNP avec attestation et vérification de workload intégrées.



Technologies TEE Inférence confidentielle Entraînement multi-parties



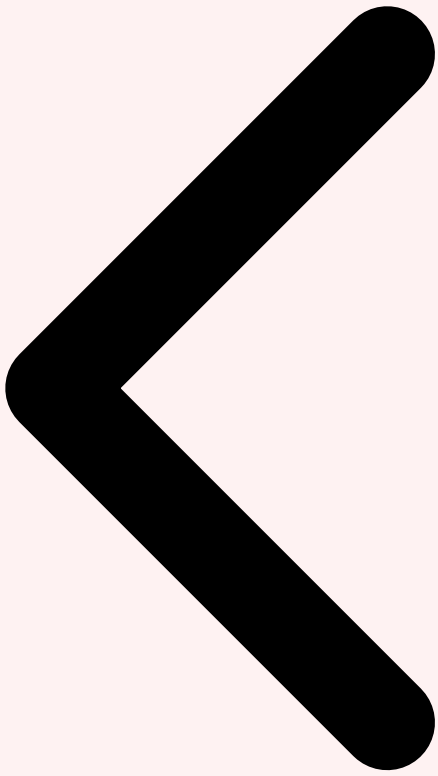
Comment garantir que vos modèles de machine learning ne deviennent pas des vecteurs d'attaque ?

4 Entraînement multi-parties sécurisé

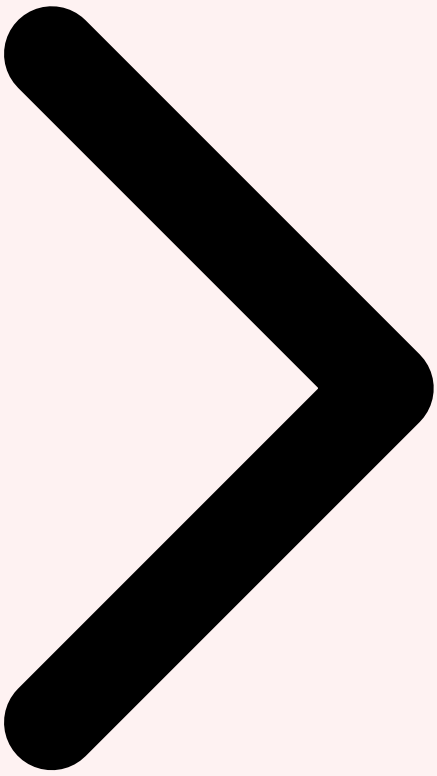
L'**entraînement multi-parties sécurisé** permet à plusieurs organisations de contribuer leurs données à l'entraînement d'un modèle commun sans que les données de chaque partie ne soient exposées aux autres. Ce cas d'usage répond à un besoin critique : dans de nombreux domaines (santé, finance, défense), les données d'entraînement sont réparties entre plusieurs organisations qui ne peuvent pas les partager pour des raisons réglementaires ou concurrentielles, mais qui bénéficieraient d'un modèle entraîné sur l'ensemble des données.

Le Confidential Computing offre une approche complémentaire au **federated learning** pour résoudre ce problème. Dans le federated learning, chaque partie entraîne localement et ne partage que les gradients — mais les gradients peuvent leaker des informations sur les données d'entraînement (gradient inversion attacks). Avec le Confidential Computing, les données brutes de chaque partie sont chargées dans un TEE centralisé où

l'entraînement complet est exécuté de manière confidentielle. Aucune partie ne peut accéder aux données des autres, et l'opérateur de l'infrastructure ne peut accéder à aucune donnée. Le modèle résultant est extrait du TEE selon des règles de gouvernance prédéfinies (par exemple, seuls les poids du modèle sortent, pas les données). Des projets comme **Cape Privacy**, **Opaque Systems** et le consortium **MELLODDY** (pharmaceutique) implémentent cette approche en production. Pour approfondir, consultez Comment Choisir sa Base.



Inférence Entraînement multi-parties Attestation

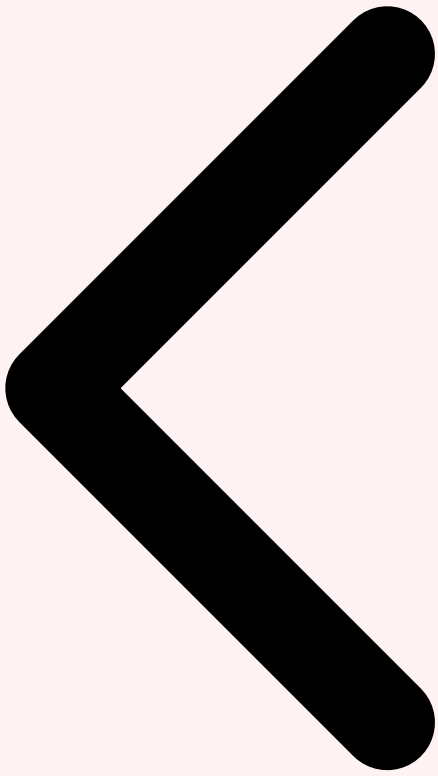


Cas concret
En février 2024, une entreprise de Hong Kong a perdu 25 millions de dollars après qu'un employé a été trompé par un deepfake vidéo lors d'une visioconférence. Les attaquants avaient recréé l'apparence et la voix du directeur financier à l'aide de modèles d'IA générative, démontrant les risques concrets de cette technologie en contexte corporate.

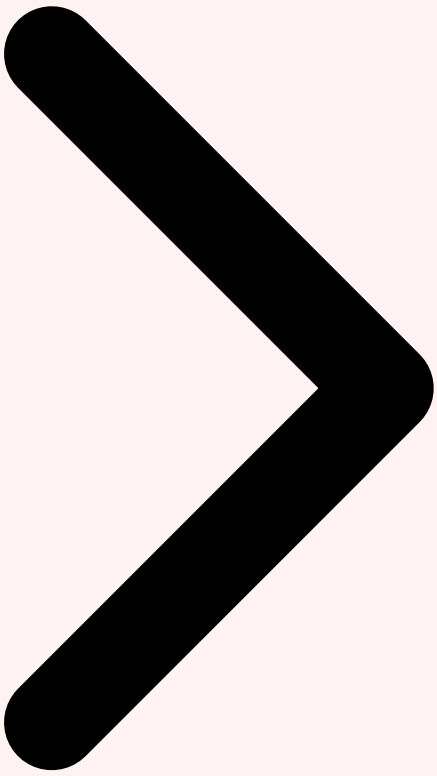
5 Attestation de modèles

L'**attestation** est le mécanisme par lequel un TEE prouve cryptographiquement à un tiers que le code et la configuration exécutés correspondent à ce qui est attendu. Dans le contexte IA, l'attestation permet de vérifier que le modèle déployé est bien celui qui a été audité, que le moteur d'inférence n'a pas été modifié, et que l'environnement d'exécution est confidentiel et intègre. Le processus d'attestation génère un **rapport d'attestation** signé par le matériel (TPM, Secure Processor) contenant des mesures cryptographiques (hashes) du code, de la configuration et de l'état initial du TEE.

L'**attestation de modèle** étend ce concept en incluant le hash des poids du modèle dans le rapport d'attestation. Un utilisateur peut ainsi vérifier, avant d'envoyer ses données, que le modèle exact qui traitera sa requête est un modèle spécifique et audité, et non une version modifiée (par exemple backdoorée). Les services d'attestation comme **Microsoft Azure Attestation (MAA)**, **Intel Trust Authority** et **Google Confidential Space Attestation** fournissent des API pour vérifier les rapports d'attestation. Le protocole **RATS (Remote ATtestation procedureS)** de l'IETF standardise les formats et les flux d'attestation pour garantir l'interopérabilité entre les implémentations TEE.



Entraînement Attestation Azure Confidential

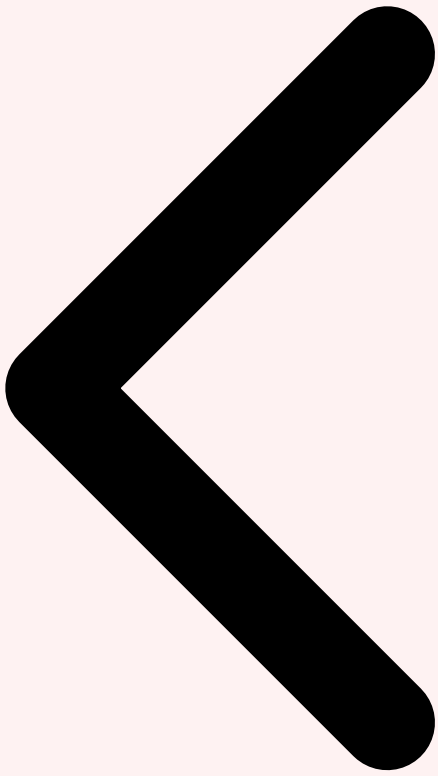


6 Azure Confidential Computing + IA

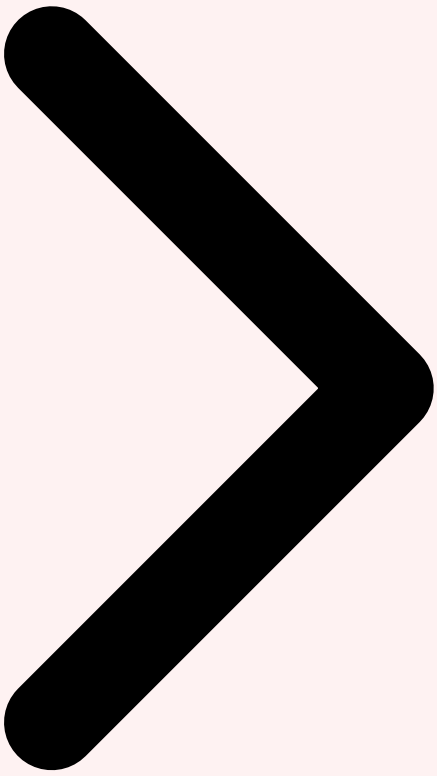
Microsoft Azure est le cloud provider le plus avancé en matière de Confidential Computing pour l'IA, avec une offre complète couvrant l'inférence, le fine-tuning et l'entraînement. Les **VMs confidentielles DCsv3/DCdsv3** basées sur AMD SEV-SNP offrent jusqu'à 96 vCPUs et 384 Go de mémoire protégée pour les workloads IA CPU-only. Les **VMs confidentielles avec GPU** (NCCsv3 avec NVIDIA H100) permettent l'inférence et le fine-tuning de modèles avec des garanties de confidentialité sur le CPU et le GPU. **Azure Confidential Ledger** fournit un registre immuable pour l'audit des opérations de déploiement et d'attestation des modèles.

Azure OpenAI Service avec Confidential Inference permet d'utiliser les modèles GPT-4o et GPT-4 Turbo dans un environnement confidentiel où Microsoft ne peut pas accéder aux prompts ni aux réponses. **Azure Machine Learning Confidential** intègre les VMs confidentielles dans les pipelines AzureML, permettant le fine-tuning de modèles sur des données sensibles sans exposition au cloud provider. Le **Azure Confidential Clean Room** fournit un environnement multi-parties sécurisé pour l'entraînement collaboratif, avec des

politiques de gouvernance définies par les participants et appliquées par le matériel. En complément, **GCP Confidential Space** et **AWS Nitro Enclaves** offrent des capacités comparables sur leurs plateformes respectives.



Attestation Azure Confidential Performances



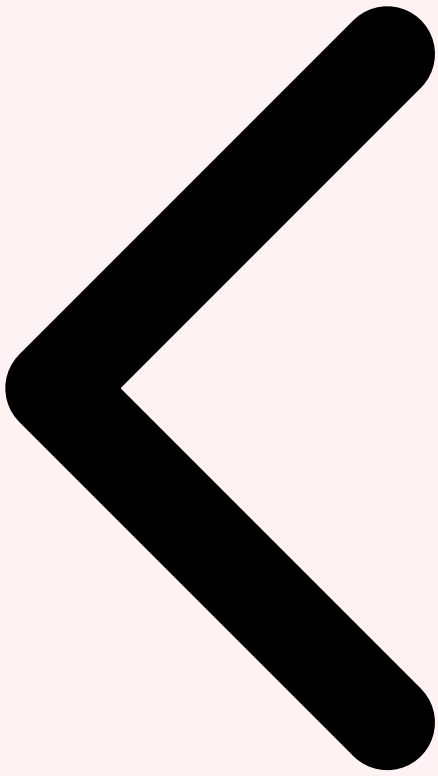
7 Performances et overhead

L'**overhead de performance** du Confidential Computing est un facteur critique pour les workloads IA, particulièrement sensibles à la latence et au throughput. Les technologies VM-level (TDX, SEV-SNP) offrent un overhead significativement plus faible que les technologies application-level (SGX). Pour AMD SEV-SNP, le chiffrement mémoire AES-256 est effectué par le contrôleur mémoire en matériel, avec un overhead typique de **2 à 5%** sur les workloads compute-intensive comme l'inférence ML. Intel TDX présente un profil similaire. L'impact principal provient des transitions entre le monde confidentiel et le monde hôte (VM exits), qui sont plus fréquentes pour les workloads I/O-intensive que pour les workloads compute-intensive.

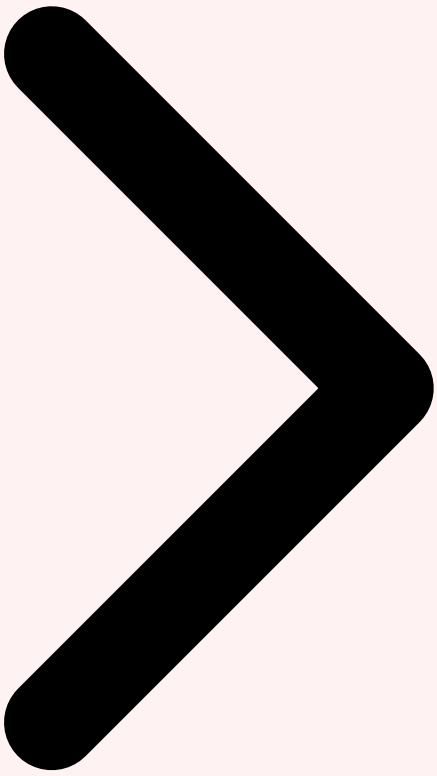
Pour les **GPU confidentiels** NVIDIA H100, l'overhead provient du chiffrement du bus PCIe entre le CPU et le GPU et du chiffrement de la mémoire HBM. Les benchmarks publiés par NVIDIA indiquent un overhead de **5 à 10%** sur les workloads d'inférence LLM, et de 10 à 15% sur l'entraînement. L'impact est plus prononcé pour les modèles nécessitant des échanges fréquents entre CPU et GPU (embedding lookups, preprocessing). Pour les

modèles dont le compute est dominé par les opérations matricielles sur GPU (attention, FFN), l'overhead est minimal. Les optimisations continues du firmware et des drivers NVIDIA réduisent progressivement cet overhead. Pour approfondir, consultez IA Multimodale : Texte, Image et Audio.

En termes de **coût**, les VMs confidentielles sont typiquement 10 à 20% plus chères que les VMs standard équivalentes, reflétant le coût du matériel TEE et de l'attestation. Pour les workloads IA où la confidentialité est non négociable (santé, finance, défense), ce surcoût est largement justifié par rapport aux alternatives (on-premise dédié, chiffrement homomorphe avec un overhead de 1000x, ou renoncement au cloud). Le calcul TCO doit intégrer les économies en matière de compliance (RGPD, HIPAA, AI Act) et de gestion des risques de fuite de données.



Azure Confidential Performances Conclusion



8 Conclusion et perspectives

Le **Confidential Computing** transforme fondamentalement la posture de sécurité des déploiements IA en éliminant la nécessité de faire confiance à l'opérateur de l'infrastructure. Les technologies TEE (Intel TDX, AMD SEV-SNP, ARM CCA) combinées aux GPU confidentiels NVIDIA permettent désormais l'inférence, le fine-tuning et l'entraînement de modèles IA avec des garanties de confidentialité vérifiables par attestation cryptographique, et un overhead de performance acceptable (2-15% selon le workload).

Les cas d'usage les plus immédiats sont l'inférence confidentielle de LLM sur des données sensibles (santé, juridique, finance), l'entraînement multi-parties dans les consortiums industriels et de recherche, et la protection de la propriété intellectuelle des modèles dans les déploiements cloud. Les offres cloud (Azure Confidential AI, GCP Confidential Space, AWS Nitro Enclaves) rendent ces capacités accessibles sans expertise matérielle spécifique.

Les perspectives incluent le Confidential Computing homomorphe (combinaison TEE + HE pour une protection en couches), les GPU confidentiels de prochaine génération avec un overhead réduit, et l'attestation continue des pipelines MLOps complets.

Recommandations : Si vos modèles IA traitent des données sensibles dans le cloud, évaluez dès maintenant les offres de Confidential Computing. Commencez par l'inférence confidentielle (le cas d'usage le plus mature), intégrez l'attestation dans vos workflows de déploiement, et planifiez la migration des workloads de fine-tuning vers des VMs confidentielles avec GPU. Le surcoût de 10-20% est un investissement négligeable face aux risques de fuite de données et de non-conformité réglementaire.

Besoin d'un accompagnement expert ?

Nos consultants vous accompagnent dans la mise en place d'architectures IA confidentielles et l'intégration du Confidential Computing dans vos pipelines MLOps. Devis personnalisé sous 24h.

Références et ressources externes

- OWASP LLM Top 10 — Les 10 risques majeurs pour les applications LLM
- MITRE ATLAS — Framework de menaces pour les systèmes d'intelligence artificielle
- NIST AI RMF — AI Risk Management Framework du NIST
- arXiv — Archive ouverte de publications scientifiques en IA
- HuggingFace Docs — Documentation de référence pour les modèles de ML

Pour approfondir ce sujet, consultez notre outil open-source llm-vulnerability-scanner qui facilite l'analyse des vulnérabilités des LLM.

Sources et références : ArXiv IA · Hugging Face Papers

FAQ

Qu'est-ce que Confidential Computing et IA ?

Le concept de Confidential Computing et IA 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 Confidential Computing et IA est-il important en cybersécurité ?

La compréhension de Confidential Computing et IA permet aux équipes de sécurité d'améliorer leur posture défensive. Les sections « Table des Matières » et « 1 Introduction au Confidential Computing pour l'IA » 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 Introduction au Confidential Computing pour l'IA, 2 Technologies TEE (Intel TDX, AMD SEV-SNP, ARM CCA). 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