# L'Inférence "Zéro-Connaissance" — Quand l'IA décide sans voir

2 June 2026 • 22 min de lecture

ZeroKnowledge zkML PrivacyPreservingAI FHE MPC EUAIAct RGPD ConfidentialAI ResponsibleAI


Un modèle parfaitement attesté, impeccablement calibré, reste un modèle qui "voit" ce qu'on lui soumet. Pour la santé, la finance, le droit (les domaines à plus forte valeur ajoutée et les plus rigoureusement régulés), cette visibilité suffit à bloquer l'usage : tant que la donnée circule en clair jusqu'au modèle, la confiance repose sur un contrat, pas sur une preuve.

Les preuves à divulgation nulle de connaissance (ZKP), combinées au chiffrement homomorphe (FHE) et au calcul multipartite sécurisé (MPC), proposent une réponse structurellement différente : retirer la nécessité même de voir la donnée pour en certifier une conclusion. L'IA cesse d'être un expert consulté — elle devient un expert aveugle dont la parole est mathématiquement opposable.

## Le verrou que l'attestation ne résout pas

L'attestation cryptographique prouve que le modèle exécuté est bien celui déclaré. Elle ne change rien au fait que ce modèle, pendant l'inférence, traite la donnée en clair. Or tout traitement de données personnelles au sens du RGPD, y compris leur simple lecture par un système tiers, exige une base légale valide (Art. 6), une finalité explicite (Art. 5(1)(b)), et, pour les catégories sensibles, une protection renforcée (Art. 9).

Anonymiser des données avant de les soumettre à un LLM cloud ne supprime pas le risque, elle le déplace. L'étude de référence de Rocher et al. (*Nature Communications*, 2019) a démontré que 99,98 % des individus sont ré-identifiables dans un jeu de données anonymisé à partir de seulement 15 attributs démographiques courants. La conclusion des auteurs est sans appel : le pseudonymat, la k-anonymisation, la suppression de champs — ces techniques réduisent le risque statistiquement, elles ne l'éliminent pas (Rocher et al., *Estimating the success of re-identifications in incomplete datasets using generative models*, Nature Communications, 2019). Le RGPD exige une élimination, pas une réduction. Les orientations du Comité Européen de la Protection des Données (CEPD) de 2023 sur l'IA confirment cette lecture : l'anonymisation insuffisante dans un contexte d'inférence IA est un traitement illicite, pas une précaution insuffisante.

Le résultat est un marché structurellement verrouillé : les cas d'usage les plus prometteurs de l'IA (diagnostic différentiel sur dossier patient, scoring de crédit sur relevés bancaires, analyse de conformité sur contrats sous NDA, triage médical d'urgence ...) sont précisément ceux que la contrainte d'exposition des données rend juridiquement impraticables dans une architecture cloud standard.

# L'expert aveugle — une propriété mathématique, pas une métaphore

## Une analogie pour ancrer l'intuition

Imaginez Harry Potter devant le coffre de Godric Gryffondor, dans les profondeurs de Gringotts. La serrure exige qu'il prouve qu'il connaît le code secret — mais s'il révèle le code au gobelin Griphook, celui-ci pourra revenir le voler. Harry veut donc **prouver qu'il connaît la solution sans jamais la révéler**. C'est exactement ce que fait une preuve à divulgation nulle de connaissance (ZKP) : fournir une certitude mathématique sur un fait, sans exposer les données qui fondent ce fait.

En restant dans l'univers de J.K Rowling, nous pouvons aussi imaginer que Madame Pomfrey doit certifier qu'un patient a bien été exposé à un sortilège de la liste des maléfices interdits, sans jamais accéder à son dossier médical complet. Elle ne peut pas lire le dossier, la loi de protection des secrets magiques le lui interdit. Mais elle peut vérifier une empreinte cryptographique : une preuve que les biomarqueurs du patient tombent bien dans la plage associée aux maléfices de la liste, sans que cette plage révèle aucun des biomarqueurs eux-mêmes.

C'est précisément la propriété que ZKP confère à l'inférence IA : le modèle peut certifier une conclusion sur une donnée sans que cette donnée ne lui soit jamais accessible en clair.

### Les trois rôles du protocole

Dans toute architecture d'inférence ZK, trois rôles sont structurellement distincts :

| Rôle | Qui | Ce qu'il détient | Ce qu'il apprend |
|---|---|---|---|
| **Prouveur** | Le patient / l'emprunteur / le client | La donnée brute | Il génère et contrôle la preuve |
| **Modèle aveugle** | L'IA du fournisseur | Le circuit de calcul (poids) | Rien sur l'input |
| **Vérificateur** | L'assureur / le régulateur / le juge | La preuve et la conclusion | La conclusion uniquement |

Le différenciateur fondamental par rapport à toute autre approche de confidentialité : l'impossibilité n'est pas contractuelle, elle est arithmétique. Le fournisseur ne s'engage pas à ne pas regarder la donnée, il est dans l'impossibilité technique de le faire. Aucun NDA, aucune clause de sous-traitance, aucune politique de rétention ne peut offrir cette garantie.

---

## La carte des techniques — ZKP, FHE, MPC

Trois familles de cryptographie répondent à ce besoin, avec des compromis distincts. Aucune n'est universellement supérieure ; le choix dépend du modèle cible, de la contrainte de latence et du niveau de garantie requis.

**ZKP / zkML — la preuve sans révélation.** Le Prouveur calcule localement et soumet une preuve mathématique de la conclusion. Le fournisseur vérifie la preuve sans jamais accéder à l'input. Overhead en 2026 : de quelques secondes sur des circuits arithmétiques minimaux (preuve de seuil sur un score scalaire) à plusieurs minutes pour des classificateurs légers de type Random Forest ou réseau convolutif ~18M paramètres, jusqu'à des dizaines de minutes pour GPT-2 (avec NANOZK, 43 s ; avec EZKL baseline, plusieurs heures) — les LLMs à plusieurs milliards de paramètres restent hors de portée en mode non-asynchrone. Force principale : la vérifiabilité est cryptographiquement parfaite. Limite : le coût de preuve croît avec la taille du circuit, pas seulement avec celle du modèle.

**FHE — le calcul sur donnée chiffrée.** La donnée ne quitte jamais son état chiffré. Le fournisseur exécute le modèle directement sur le chiffré et renvoie un résultat chiffré que seul le Prouveur peut déchiffrer. La garantie est absolue : le fournisseur n'a physiquement aucun accès à quoi que ce soit d'intelligible. Overhead : ×100 à ×1 000 en 2026 pour les réseaux légers, prohibitif pour les LLMs. Trajectoire : les progrès des schémas CKKS et TFHE réduisent régulièrement cet écart.

**MPC — le calcul distribué sans connaissance individuelle.** Le calcul est réparti entre plusieurs nœuds de telle sorte qu'aucun d'eux, individuellement, ne dispose de l'information complète. La sécurité est garantie tant qu'un sous-ensemble de nœuds reste honnête. Overhead : ×10 à ×50 selon la latence réseau. Contrainte : requiert une infrastructure entre les parties calculantes — adapté aux consortiums inter-entreprises, moins aux architectures client-serveur classiques.

| Technique | Overhead 2026 | Modèle cible réaliste | Donnée cachée à | Garantie | Maturit produc |
|---|---|---|---|---|---|
| **ZKP / zkML** | Secondes (seuil/régression) → minutes (classif. légère) → ~15 min/inférence (13B, zkLLM) | Scoring/seuil (prod.), classification légère (pilote), LLMs >1B (recherche) | Vérificateur | Cryptographique | Partiel (EZKL[1], — mod légers uniquer |
| **FHE** | ×100–1 000 | ML traditionnel, CNN léger | Fournisseur | Absolue | POC / product limitée |

| Technique | Overhead 2026 | Modèle cible réaliste | Donnée cachée à | Garantie | Maturit produc |
|---|---|---|---|---|---|
| **MPC** | ×10–50 | PyTorch général | Chaque partie | Seuil d'honnêteté | Oui (CrypTe SecretF |
| **ZKP + FHE hybride** | Élevé | Modèles légers ciblés | Fournisseur + Vérificateur | Maximale | Recher (TRL 3- |

> [1] ***EZKL*** *(Easy Zero-Knowledge Learning) est un framework open source qui compile des modèles ML standards (ONNX, PyTorch) en circuits arithmétiques ZK, permettant de générer et vérifier des preuves d'inférence sans réécrire le modèle. Il constitue en 2026 l'outil de référence pour les équipes qui déploient des preuves de classification ou de seuil en production. (ezkl.xyz)*

# Trois cas d'usage : du concept à la facture

## Diagnostic médical "zéro-accès"

**Le contexte réglementaire.** Les données de santé constituent une catégorie spéciale au sens de l'Art. 9 du RGPD. Leur traitement est interdit sauf base légale renforcée — Art. 9(2)(h) pour les soins de santé directs, consentement explicite pour d'autres finalités. L'European Health Data Space (EHDS), dont le déploiement s'accélère en 2026, renforce ces exigences tout en créant un cadre pour le partage secondaire — mais uniquement sous conditions strictes d'accès contrôlé. En pratique, la majorité des cas d'usage IA de diagnostic secondaire (aide à la décision sur dossiers historiques, recherche épidémiologique) reste bloquée par l'incapacité à garantir la minimisation dans une architecture cloud standard.

**L'architecture ZKP.** Le patient ou son établissement génère une preuve ZK attestant que ses biomarqueurs tombent dans la plage diagnostique X, sans transmettre les valeurs en clair. Le modèle reçoit la preuve, vérifie qu'elle satisfait le circuit de diagnostic, et renvoie une conclusion certifiée. Aucune donnée de santé ne quitte le périmètre du patient au sens du RGPD. Watanabe et al. (2025) documentent une architecture pipeline exactement dans ce cas : un premier LLM infère les traits cliniques pertinents depuis le dossier patient et génère simultanément une preuve ZK ; un second LLM consomme ces traits prouvés pour produire le diagnostic différentiel — sans jamais accéder au dossier source (arXiv:2502.06425).

**Le déblocage business.** Cette architecture transforme le consentement en optionnel pour certains usages secondaires (recherche, aide au diagnostic de second recours) car la minimisation est physiquement garantie.

### Conformité contractuelle sous NDA

**Le contexte réglementaire.** Les contrats couverts par des NDA, le secret professionnel des avocats , les informations commercialement sensibles dans les due diligences M&A — autant de catégories pour lesquelles la soumission à un LLM cloud représente une divulgation au sens du droit des contrats et, pour les avocats, une violation potentielle du secret professionnel.

**L'architecture ZKP.** L'avocat ou le compliance officer génère une preuve ZK attestant que le contrat satisfait (ou non) un ensemble de critères réglementaires ( absence de clause contraire au RGPD Art. 28, présence des mentions obligatoires LCEN, conformité avec les standards ISDA) sans que le texte du contrat ne soit transmis au modèle en clair. La preuve certifie la conformité ou non-conformité sans exposer le contenu. ZKPROV (arXiv:2506.20915), conçu initialement pour la provenance des données d'entraînement, propose des mécanismes directement transposables à la certification de propriétés sur des documents sans révélation de leur contenu (arXiv:2506.20915).

**Le déblocage business.** Les cabinets d'avocats et les équipes M&A peuvent offrir des services d'analyse IA sur des documents clients sans que ces documents ne quittent jamais la maîtrise de ces clients — un argument décisif dans un marché où la confidentialité est la valeur principale du service.

## Mise en œuvre : le chemin réaliste

### Verrous ouverts spécifiques à la privacy de l'input

**Input size vs circuit size.** Plus l'input est grand — texte long, image haute résolution, dossier médical complet — plus le circuit ZK est vaste. La propriété zero-knowledge a un coût qui croît avec la taille de ce qu'on cache, pas seulement avec la complexité du modèle.

**Auditabilité sélective.** Cas réel et encore ouvert : un régulateur veut vérifier qu'une décision de crédit défavorable n'était pas discriminatoire, sans accéder aux données personnelles du demandeur. Cela requiert des preuves ZK *partielles*, prouver une propriété sur un sous-ensemble de features (le genre et l'origine n'ont pas influencé la décision) sans révéler les autres features. Problème formellement ouvert en 2026.

**Privacy du modèle vs privacy de l'input.** Les systèmes qui protègent l'input exposent souvent les poids au Prouveur ; ceux qui protègent les poids exposent l'input au fournisseur. Les architectures qui garantissent les deux simultanément avec des performances acceptables constituent le problème central du zkML privacy-oriented. C'est aussi l'intersection la plus directement pertinente pour la propriété intellectuelle des fournisseurs et la protection des données des utilisateurs.

**Composabilité des preuves.** Dans l'architecture Watanabe (arXiv:2502.06425), la preuve du premier LLM doit être vérifiable par le second sans révéler l'input original. Les preuves récursives (Nova, HyperNova) sont la réponse, mais leur intégration dans des pipelines ML multi-étapes reste un problème ouvert dont la résolution conditionne le déploiement à grande échelle de l'inférence privée en cascade.

## Ce qui est industrialisable aujourd'hui

La frontière entre le possible et le fantasme est nette en 2026. Trois patterns sont déployables sans attendre la maturité des LLMs en ZKP complet. Mais cette frontière doit être énoncée avec précision, car la littérature commerciale tend à la déplacer vers l'optimisme.

### Pattern 1 — Preuve de seuil

Le cas le plus simple et le plus directement utile : prouver qu'une valeur scalaire (un score, un ratio, un montant) est au-dessus ou en-dessous d'un seuil sans révéler la valeur. La performance tient à la stratégie : au lieu de prouver l'inférence d'un modèle ML complet, on ne prouve qu'une comparaison arithmétique ( `score ≤ seuil` ) sur un circuit minimaliste — quelques dizaines de contraintes contre des millions pour un réseau de neurones. Les benchmarks EZKL sur des opérations de ce type (régression linéaire, SVM simple) montrent des temps de preuve de 0,1 à 6 secondes sur CPU standard selon la complexité du modèle sous-jacent (EZKL Blog, *Benchmarking ZKML Frameworks*, janvier 2024). La vérification côté serveur prend quelques millisecondes. Applicable immédiatement au scoring de crédit (seuil d'octroi), au triage médical (seuil d'alerte), à la détection de fraude binaire. **Condition nécessaire : le modèle produceur du score doit lui-même être de taille modeste** (régression, arbre de décision, SVM) — la preuve de seuil ne s'applique pas en aval d'un LLM.

### Pattern 2 — Preuve de classification

Prouver qu'un document appartient à une catégorie réglementaire sans exposer le document. La stratégie ici est la **quantification** : les modèles ML opèrent en virgule flottante, mais les circuits ZK n'acceptent que des entiers en corps finis. EZKL convertit automatiquement les poids et activations en arithmétique entière à précision fixe avant compilation en circuit — avec une perte de précision typiquement inférieure à 2 points sur les tâches de classification standard. En revanche, les temps de preuve pour des modèles de type BERT-small (66M paramètres) restent de l'ordre de **plusieurs minutes sur GPU** en 2026 — des benchmarks indépendants (Emergent Mind, *Topics: EZKL*, 2025) signalent qu'EZKL peut nécessiter "des minutes à des heures". Ce pattern est en pilote production dans des contextes où la latence de preuve est acceptable, pas encore déployable à grande échelle.

**Pattern 3 — Inférence asynchrone avec preuve différée**

Pour les LLMs dépassant le milliard de paramètres, c'est l'honnêteté intellectuelle qui s'impose : **l'état de l'art en 2026 ne permet pas de générer des preuves ZK complètes d'inférence LLM à coût et délai industriellement acceptables.** Le meilleur résultat académique documenté — zkLLM (Sun et al., arXiv:2404.16109) — atteint moins de 15 minutes pour un modèle de 13 milliards de paramètres, sur du matériel GPU dédié, avec une architecture de preuve (sumcheck protocol + tlookup) spécifiquement conçue pour ce cas et distincte d'EZKL. C'est un record, pas une baseline déployable.

Le pattern asynchrone consiste à dissocier la contrainte métier (décision temps-réel) de la contrainte réglementaire (preuve archivée pour audit). La logique est recevable — la majorité des obligations réglementaires requièrent une preuve disponible à l'audit, pas une preuve instantanée. Mais ses limites de scalabilité sont sévères : à raison de 15 minutes par preuve sur un GPU dédié, 1 000 requêtes quotidiennes représentent déjà 250 GPU-heures, soit un coût d'infrastructure considérable avant même d'aborder le backlog en cas de pic. Ce pattern est crédible uniquement pour des flux régulés à **volume faible et valeur unitaire élevée** (décisions de crédit corporate, diagnostics de second recours sur dossiers complexes) — pas pour des pipelines à haute fréquence.

Les stratégies alternatives activement explorées pour contourner ces limites sont : la **vérification par échantillonnage** (VeriLLM : environ 1 % du coût d'inférence, mais garantie probabiliste et non cryptographique), les **empreintes heuristiques** (TOPLOC : rapide mais non opposable devant un régulateur), et le **fine-tuning vérifiable par LoRA** (zkLoRA : prouve les

adaptateurs, pas le modèle de base). Aucune n'offre la garantie cryptographique complète de l'inférence ZK. Ce sont des compromis pragmatiques, pas des solutions équivalentes.

## L'honnêteté sur l'état de l'art : les LLMs >1B paramètres sont un problème ouvert

Il convient d'énoncer clairement ce que la littérature académique dit, sans l'habiller en solution industrielle prête à l'emploi.

**Le plafond documenté.** En 2026, le meilleur système académique de preuve ZK complète pour un LLM est zkLLM (Sun et al., arXiv:2404.16109) : moins de 15 minutes pour un modèle LLaMA-2 à 13 milliards de paramètres, sur GPU dédié, avec une architecture sumcheck + tlookup spécialement conçue pour les transformers. C'est 50× plus rapide que les approches génériques précédentes — et c'est encore 15 minutes. NANOZK (arXiv:2603.18046, mars 2026) descend à 43 secondes sur GPT-2 (117M paramètres) grâce à la décomposition par couches : un progrès remarquable, qui reste sur un modèle deux ordres de grandeur plus petit que les LLMs en production. Les LLMs dépassant 7 milliards de paramètres restent hors de portée d'une preuve ZK complète à coût acceptable en 2026. Ce n'est pas une limite conjoncturelle : elle est structurelle, liée à la croissance du circuit arithmétique avec la taille du modèle.

**Pourquoi c'est difficile — les trois verrous fondamentaux.** Premièrement, la *taille du circuit* : un seul passage avant sur GPT-2 implique plus de 100 millions de multiplications ; représenter cela en circuit arithmétique produit des milliards de contraintes. Deuxièmement, les *opérations non-arithmétiques* : softmax, GELU, LayerNorm ne s'expriment pas naturellement en corps finis — leur arithmétisation approximative (lookup tables, tlookup) introduit soit une perte de précision soit un overhead de preuves supplémentaires. Troisièmement, la *mémoire* : générer la preuve d'un LLM de 13B paramètres nécessite de maintenir en RAM l'intégralité des tenseurs intermédiaires pour la construction de la preuve, ce qui dépasse les capacités des systèmes génériques et explique les erreurs out-of-memory observées sur zkML pour GPT-2 1.5B.

**Quelle est alors la meilleure stratégie disponible, selon le contexte ?**

*En temps réel strict (latence < quelques secondes, LLM >1B paramètres).* Il n'existe pas en 2026 de preuve ZK complète industrialisable dans ce cadre. Les alternatives disponibles sont, par ordre de garantie décroissante : (a)

**TOPLOC** — empreinte polynomiale légère des activations intermédiaires (258 octets par 32 tokens), vérification plus rapide que l'inférence originale, détection fiable des modifications du modèle, mais ce n'est pas une preuve ZK : c'est une empreinte heuristique. Elle est opposable sur le plan technique mais pas cryptographiquement au sens formel, et sa valeur devant un régulateur reste à établir par la jurisprudence. (b) **VeriLLM** — re-vérification empirique d'environ 1 % des calculs, architecture d'incitation économique dissuadant la triche dans les réseaux décentralisés, mais garantie probabiliste. (c) **Architecture d'isolation par modèle léger** : ne pas tenter de prouver l'inférence du LLM lui-même, mais faire traiter l'input par un classificateur léger *avant* le LLM — seul ce classificateur léger est soumis à preuve ZK (Pattern 1 ou 2 ci-dessus). Le LLM reçoit des features certifiées, pas l'input brut. Cette architecture ne prouve pas ce que le LLM *fait* de ces features, mais elle garantit qu'il ne voit pas l'input brut — ce qui peut suffire pour certaines obligations de minimisation RGPD.

*En temps différé (audit asynchrone, LLM >1B paramètres).* Le Pattern 3 (preuve générée en batch différé) reste la voie la plus réaliste, sous réserve de son dimensionnement honnête : il est crédible pour des flux à volume faible et valeur unitaire élevée, pas pour des pipelines haute fréquence. Pour améliorer sa scalabilité, deux pistes complémentaires sont actives : la **décomposition par couches** (stratégie NANOZK — prouver chaque couche indépendamment et en parallèle, ce qui permet une parallélisation horizontale sur plusieurs GPU) et le **fine-tuning vérifiable** (zkLoRA — prouver que les adaptateurs LoRA ont été appliqués correctement au modèle de base, sans prouver le modèle de base lui-même). Aucune de ces pistes ne supprime le coût fondamental ; elles le répartissent et le parallélisent.

**La recommandation honnête pour l'industrie en 2026.** Pour tout cas d'usage impliquant un LLM de plus d'un milliard de paramètres, la preuve ZK complète de l'inférence n'est pas une option industrielle aujourd'hui. La stratégie réaliste est une combinaison : prouver cryptographiquement ce qui peut l'être (les features d'entrée, le franchissement d'un seuil, la classification préalable) et recourir pour le reste à des architectures complémentaires — TEE (Trusted Execution Environment) sur GPU confidentiel pour l'isolation d'exécution, journalisation immuable, MPC pour la distribution du calcul entre parties de confiance partielle. Ce n'est pas une défaite : c'est la même trajectoire que TLS dans les années 2000, où les premières implémentations coûtaient des ordres de grandeur d'overhead avant de devenir transparentes. La date d'inflexion pour les LLMs lourds n'est pas 2026 — mais les organisations qui

construisent les architectures modulaires aujourd'hui seront prêtes à activer la couche ZK dès qu'elle deviendra disponible à coût acceptable.

### Matrice de criticité et stack recommandée

| Criticité | Cas typique | Stack privacy input |
|---|---|---|
| **Faible** | Chatbot interne, résumé non-sensible | Pseudonymat + politique de traitement |
| **Moyenne** | RAG sur données RH ou clients B2B | EZKL preuve de classification sur modèle léger (<100M params) + MPC CrypTen |
| **Haute** | Scoring crédit (modèle léger), diagnostic seuil, NDA | Concrete ML FHE ou EZKL preuve de seuil arithmétique |
| **Haute — LLM >1B** | Diagnostic LLM, analyse contractuelle LLM | TEE GPU confidentiel + isolation features par classificateur ZK amont + audit asynchrone |
| **Extrême** | Dossiers défense, données de santé Art. 9 | ZKP + FHE hybride + TEE GPU confidentiel |

**Délai réaliste** : un POC "preuve de seuil en production" (circuit minimal sur modèle léger, intégration, audit partiel) demande 2 à 3 mois pour une équipe de 2 ingénieurs ZK et 1 ingénieur ML. Une intégration complète sur un pipeline de classification légère (<100M paramètres) : 4 à 6 mois, 3 à 4 personnes. Pour un LLM >1B avec preuve asynchrone différée : 8 à 18 mois minimum, équipe spécialisée zkML, avec des garanties cryptographiques partielles — et ce délai suppose que la technologie ne dépasse pas l'état de l'art 2026.

## Objections et réponses factuelles

**« C'est trop lent pour la production. »** La réponse honnête est nuancée. Pour les preuves de seuil arithmétique (circuit minimal sur un score scalaire issu d'un modèle léger — régression, SVM), les benchmarks EZKL montrent des temps de 0,1 à 6 secondes sur CPU, vérification en millisecondes : ce pattern est opérationnel aujourd'hui. Pour la preuve de classification sur des modèles de type BERT-small, les temps sont de l'ordre de plusieurs minutes — acceptable pour un usage asynchrone, pas pour du temps-réel. Pour les LLMs dépassant le milliard de paramètres, l'état de l'art (zkLLM, 13B paramètres) est à moins de 15 minutes sur GPU dédié — ce qui n'est ni du

temps-réel, ni scalable à volume. La preuve asynchrone décorrèle utilement la décision temps-réel de la preuve réglementaire pour des flux à volume modéré. Elle ne résout pas la question pour les pipelines à haute fréquence, où les alternatives probabilistes (VeriLLM, TOPLOC) restent les seules options industrialisables en 2026 — avec des garanties cryptographiques moindres.

**« On peut anonymiser les données — c'est plus simple. »** Non. L'étude Rocher et al. a démontré que 99,98 % des individus sont ré-identifiables dans un jeu de données "anonymisé" à partir de 15 attributs démographiques courants (Nature Communications, 2019). L'anonymisation RGPD est une garantie statistique, pas mathématique. ZKP est une garantie mathématique. La distinction est juridiquement décisive devant une DPA — et les orientations du CEPD depuis 2023 la formalisent de façon croissante.

**« Le régulateur ne comprend pas encore. »** La CNIL a publié ses travaux sur les PETs dès 2023, incluant ZKP et FHE parmi les techniques recommandées pour le traitement de données de santé et financières. L'ENISA a mis à jour son catalogue PET en 2025 avec une section explicite sur l'inférence IA privée. Le GPAI Code of Practice mentionne les techniques cryptographiques de confidentialité dans ses lignes directrices techniques. Être en avance sur la courbe réglementaire, c'est contribuer à en écrire les standards — un avantage concurrentiel documenté dans les secteurs fintech et medtech.

**« FHE sur LLM est encore inaccessible pour une organisation standard. »** C'est exact pour les LLMs lourds — et ce n'est pas là que commence le déploiement. Les frameworks open source (EZKL, Concrete ML, CrypTen) ont ramené le coût d'entrée sur les cas légers (scoring, classification) à celui d'une équipe de 2-3 ingénieurs formés. La dette n'est pas le framework : c'est la formation. Et ce coût baisse régulièrement, comme ce fut le cas pour TLS dans les années 2000 — personne ne prétendait en 1998 que HTTPS était "trop complexe pour les PME".

**« On perd en précision du modèle avec la quantification FHE. »** Sur les réseaux légers, la perte de précision liée à la quantification TFHE est documentée à 1 à 3 points de précision par Zama sur les benchmarks UCI standard — dans la marge acceptable pour la plupart des applications régulées. Sur les LLMs, FHE reste impraticable ; ZKP avec input privé (pattern EZKL) ne nécessite pas de quantification du modèle, uniquement une représentation adaptée de l'input dans le circuit arithmétique.

---

## Conclusion

L'avantage compétitif qui s'ouvre n'est pas marginal. Les marchés santé, finance et droit représentent l'essentiel de la valeur IA non encore déployée. McKinsey les chiffre entre 400 et 600 milliards de dollars de valeur débloquable par les PETs sur 2025-2030. Ces marchés ne sont pas inaccessibles par manque de modèles performants — ils le sont par impossibilité de traiter les données qui les alimentent. ZKP, FHE et MPC ne créent pas de valeur intrinsèque : ils déverrouillent la valeur des cas d'usage déjà identifiés et jusqu'ici inaccessibles.

Pour les acteurs européens, éditeurs IA, intégrateurs systèmes, opérateurs d'infrastructures critiques, la convergence des régimes (RGPD, AI Act, EHDS, NIS2, DORA, ISO 42001) compose un brief stratégique cohérent avec cette direction. Les organisations qui transformeront ces exigences en compétences industrielles — architectures ZKP de production, pipelines FHE, MPC entre consortiums régulés — auront en 2027-2028 un *moat* défensif que peu de concurrents non-européens pourront répliquer rapidement. Non parce que les technologies sont secrètes, mais parce que la capacité à les intégrer dans des architectures de production conformes, auditées et certifiables est une compétence d'organisation, pas seulement de code.

L'expert aveugle n'est pas un expert diminué. C'est le seul expert admis dans les salles où les décisions comptent vraiment. Et pour la première fois, sa parole est mathématiquement opposable.

***Lego VIII - sac 3 - Substractive strategy***

*Les opinions exprimées dans cet article sont strictement personnelles et ne reflètent pas nécessairement celles de mon employeur. Les contenus sont fournis à titre informatif et ne constituent pas un conseil juridique. Cet article explore des concepts architecturaux émergents et analyse des tendances de marché.*

---



© 2026 Eric Blaudez. All rights reserved.

Les opinions exprimées sur ce site sont strictement personnelles et ne reflètent pas nécessairement celles de mon employeur. Les contenus sont fournis à titre informatif et ne constituent pas un conseil juridique.