Small Language Models 2026 : Phi-4, Gemma 3 et Edge AI

Découvrez les small language models 2026 : Phi-4, Gemma 3 et Edge AI. Comparatif technique, benchmarks, déploiement edge et sécurité pour professionnels.

Résumé exécutif

En 2026, les **small language models (SLM)** s'imposent comme une alternative stratégique aux LLM massifs pour les entreprises soucieuses de souveraineté, de coût et de latence. Phi-4 de Microsoft, Gemma 3 de Google DeepMind et leurs équivalents open-source atteignent des performances comparables à GPT-3.5 sur des benchmarks spécialisés, tout en s'exécutant sur du matériel standard voire sur des terminaux Edge. Ce guide technique analyse l'architecture, les performances, les vecteurs d'attaque et les stratégies de déploiement sécurisé des principaux SLM disponibles cette année.

Les **small language models 2026** représentent une rupture technologique majeure dans l'écosystème de l'intelligence artificielle appliquée à l'entreprise. Pendant des années, la course aux paramètres a dominé la recherche en NLP : plus un modèle était grand, plus il était performant. Cette logique atteint aujourd'hui ses limites opérationnelles — coût prohibitif des inférences GPU, dépendance aux hyperscalers américains, latence incompatible avec les applications temps réel et risques souverains sur les données sensibles. Les SLM, compris entre 1 et 14 milliards de paramètres, renversent cette équation. Phi-4 de Microsoft Research, entraîné sur des données synthétiques de haute qualité, dépasse des modèles dix fois plus grands sur les benchmarks de raisonnement. Gemma 3 de Google DeepMind apporte la multimodalité à moins

de 10B paramètres. Mistral, Qwen 2.5 et LLaMA 3.2 complètent un paysage open-source extraordinairement dense. Pour les équipes cybersécurité, les DSI et les RSSI, la question n'est plus de savoir si les SLM méritent attention, mais comment les évaluer, les déployer et les sécuriser dans des contextes de production exigeants, notamment sur des infrastructures Edge soumises aux contraintes ANSSI et NIS 2.

Écosystème des principaux small language models en 2026 : performances, architecture et cas d'usage Edge.

Qu'est-ce qu'un Small Language Model en 2026 ?

La *frontière SLM/LLM* est conventionnellement fixée à **14 milliards de paramètres** en 2026, bien que cette limite soit poreuse selon les architectures. Un SLM se définit davantage par ses contraintes opérationnelles que par son seul compte de paramètres : il doit pouvoir s'exécuter sur du matériel grand public (16 GB RAM, GPU discret ou NPU embarqué) avec une latence inférieure à la seconde en inférence locale. Cette définition exclut les modèles nécessitant un cluster A100, mais inclut des modèles de 27B lorsqu'ils bénéficient d'une quantification agressive.

⚡ **Retour terrain**

Dans une mission de conseil pour un groupe de services numériques, j'ai comparé deux approches d'IA générative pour la génération de rapports d'audit internes : l'une basée sur un LLM généraliste avec un prompt long, l'autre sur un modèle fine-tuné sur des rapports anonymisés. Le modèle fine-tuné était 40 % moins précis sur les sujets hors périmètre d'entraînement — mais produisait 0 % de formulations hors-charte sur le périmètre cœur. La spécialisation a un coût en généralité qu'il faut mesurer avant de décider.

La percée des SLM en 2026 repose sur trois innovations convergentes. Premièrement, la **qualité des données d'entraînement synthétiques**, popularisée par l'article fondateur *Textbooks Are All You Need* (arXiv:2309.05463) qui a démontré qu'un modèle de 1.3B paramètres entraîné sur des données pédagogiques soigneusement curées pouvait dépasser des modèles dix fois plus grands. Deuxièmement, les architectures **sparse mixture-of-experts (MoE)** qui activent seulement une fraction des paramètres lors de l'inférence. Troisièmement, les techniques de quantification avancées — GGUF, GPTQ, AWQ — analysées en détail dans notre article sur la quantization LLM 2026.

Pour les professionnels cybersécurité, les SLM ouvrent des perspectives inédites : analyse de logs en temps réel sur le SIEM, détection d'anomalies comportementales sur l'endpoint, assistance au pentesteur sur un terminal isolé. La capacité à exécuter un modèle capable dans un environnement air-gapped transforme fondamentalement l'outillage défensif.

Phi-4 : l'excellence de Microsoft Research

Phi-4 représente l'aboutissement de la recherche Microsoft sur les petits modèles denses. Disponible sur Hugging Face (microsoft/Phi-4) sous licence MIT, ce modèle de **14 milliards de paramètres** établit des records sur les benchmarks académiques les plus exigeants, dépassant des

modèles trois à cinq fois plus grands sur les tâches de raisonnement mathématique et de compréhension de code.

L'architecture de Phi-4 repose sur un transformeur dense classique — sans MoE — mais son avantage concurrentiel réside dans la composition du jeu d'entraînement. Microsoft Research a investi massivement dans la génération de *données synthétiques* de haute qualité : questions-réponses générées par GPT-4, exercices mathématiques graduels, extraits de code soigneusement annoté. Ce pipeline d'entraînement, appelé *phi-data*, privilégie la densité informationnelle sur le volume brut.

- **MMLU (5-shot)** : 84,8 % — supérieur à Llama 3.1 70B
- **HumanEval (pass@1)** : 82,6 % — comparable à GPT-4 Turbo
- **MATH benchmark** : 80,4 % — excellence en raisonnement formel
- **GSM8K** : 91,2 % — résolution de problèmes scolaires
- **Contexte** : 16 384 tokens par défaut, extensible via RoPE scaling

Du point de vue de la cybersécurité, Phi-4 présente un profil d'alignement robuste. Microsoft a appliqué un **fine-tuning RLHF étendu** avec des red teams internes focalisés sur les refus d'assistance à la création de malwares et la génération d'exploits. Les tests montrent cependant que des techniques de jailbreak avancées — notamment les injections de prompt encodées en base64 ou les prétextes fictifs — parviennent à contourner partiellement ces garde-fous, ce qui justifie un déploiement derrière une couche de filtrage applicatif.

Gemma 3 : la multimodalité open-source de Google

Gemma 3, accessible via Google AI for Developers, introduit la multimodalité native dans la famille des modèles open-weights Google DeepMind. Disponible en quatre tailles — 1B, 4B, 12B et 27B paramètres — Gemma 3 offre une flexibilité remarquable pour couvrir des usages allant du terminal IoT au serveur d'entreprise.

L'innovation architecturale centrale de Gemma 3 est la *fenêtre de contexte de 128 000 tokens*, rendue possible par une implémentation efficace de l'attention par blocs (sliding window attention alternée avec full attention). Cette capacité contextuelle dépasse la plupart des LLM commerciaux de 2024 et ouvre des cas d'usage documentaires complexes : analyse de rapports d'incident complets, corrélation de logs multi-sources, revue de code étendue.

La version 12B de Gemma 3 constitue le point d'équilibre optimal pour les déploiements enterprise en 2026. Quantifiée en Q4_K_M (format GGUF), elle s'exécute sur un serveur équipé d'un RTX 4090 ou d'un dual Xeon avec 48 GB RAM avec une latence de génération de 25 à 40 tokens/seconde — suffisant pour des assistants interactifs. Côté sécurité, Google DeepMind publie une **model card** détaillée incluant les évaluations d'alignement, ce qui facilite la due diligence avant déploiement.

Déploiement Gemma 3 12B avec Ollama

Pour tester Gemma 3 12B en local, la procédure avec Ollama (version 0.3+) est la suivante :

```
 # Installation et lancement
curl -fsSL https://ollama.com/install.sh | sh
ollama pull gemma3:12b

# Test de génération
ollama run gemma3:12b "Analyse ce log d'alerte IDS : [SYN flood sur port 443]"

# Serveur API REST local (port 11434)
ollama serve &
curl http://localhost:11434/api/generate \
  -d '{"model":"gemma3:12b","prompt":"Explique CVE-2024-3400","stream":false}'
```

La consommation mémoire observée en Q4_K_M est d'environ **8,5 GB VRAM** pour la version 12B, compatible avec la plupart des stations de travail professionnelles récentes.

Mistral et les SLM européens en 2026

Mistral AI, acteur européen majeur, maintient sa trajectoire avec **Mistral Small 3.1** (24B paramètres, contexte 128K) et **Codestral Mamba**, un modèle spécialisé code basé sur l'architecture SSM (State Space Model). Ces modèles présentent un intérêt particulier pour les organisations soumises à des contraintes de souveraineté européenne, car Mistral publie ses poids sous licence Apache 2.0 sans restriction géographique.

L'architecture *Mamba*, utilisée dans Codestral, remplace le mécanisme d'attention quadratique par une récurrence linéaire, ce qui réduit drastiquement la consommation mémoire lors du traitement de longs contextes. Sur des tâches d'analyse de codebase entières (100K+ tokens), Codestral Mamba consomme **60 % moins de mémoire** que les équivalents transformeurs, un avantage décisif pour les SIEM analysant de larges volumes de code malveillant.

En parallèle, **Qwen 2.5-Coder 14B** d'Alibaba Cloud et **DeepSeek-Coder-V2 Lite** complètent le panorama des SLM spécialisés code disponibles en 2026. Ces modèles, bien que d'origine non-européenne, sont open-weights et peuvent être audités avant déploiement — une exigence fréquente dans les politiques de sécurité des grands comptes.

Edge AI : déploiement des SLM sur terminaux contraints

Le déploiement *Edge AI* des SLM constitue l'un des changements les plus structurants de 2026 pour les équipes opérationnelles. L'Edge désigne ici tout dispositif exécutant l'inférence localement — stations de travail, serveurs de SOC, pare-feux nouvelle génération, caméras intelligentes, équipements industriels OT. Notre analyse sur l'Edge AI neuromorphique 2026 couvre les architectures matérielles en détail.

Les **NPU (Neural Processing Units)** intégrés dans les processeurs Intel Core Ultra, AMD Ryzen AI et Apple M-series atteignent en 2026 des performances suffisantes pour exécuter des SLM de 3-4B paramètres en temps réel sans GPU dédié. L'Intel Core Ultra 200V affiche 48 TOPS (tera-

opérations par seconde) sur son NPU, permettant l'exécution de Phi-4 Mini (3.8B) à 15-20 tokens/seconde en inférence INT4.

- **Format GGUF** : standard de facto pour l'exécution CPU/NPU via llama.cpp
- **ONNX Runtime** : pipeline optimisé pour les NPU Intel et AMD
- **ExecuTorch** (Meta) : framework léger pour iOS/Android et dispositifs ARM
- **CoreML** : accélération Apple Silicon avec Gemma 3 1B en 2 ms/token
- **MediaPipe LLM Inference** : Google, optimisé Android/Chrome

Pour les professionnels SOC, l'architecture cible recommandée en 2026 est un **mini-serveur edge** (NUC Pro, ASUS ExpertCenter) équipé d'un Intel Core Ultra 9, 64 GB RAM et un SSD NVMe rapide. Ce profil matériel, disponible sous 2 000 €, permet d'exécuter Gemma 3 4B ou Phi-4 Mini en continu pour l'analyse de logs sans aucune dépendance réseau vers un hyperscaler.

Benchmarks et comparatif des SLM 2026

La comparaison objective des SLM nécessite de distinguer les **benchmarks académiques généraux** (MMLU, HellaSwag, ARC) des **benchmarks applicatifs métier** (CyberSecEval, HumanEval, MATH). Les premiers mesurent la connaissance encyclopédique, les seconds la capacité opérationnelle réelle.

Modèle	Params	MMLU	HumanEval	MATH	Contexte	Licence
Phi-4	14B	84,8 %	82,6 %	80,4 %	16K	MIT
Gemma 3 12B	12B	83,2 %	78,4 %	75,1 %	128K	Gemma
Mistral Small 3.1	24B	82,7 %	79,1 %	71,3 %	128K	Apache 2.0
Llama 3.2 11B	11B	80,1 %	72,3 %	68,9 %	128K	Llama 3
Qwen 2.5 14B	14B	82,9 %	80,7 %	79,8 %	128K	Apache 2.0
Gemma 3 27B	27B	87,3 %	84,2 %	83,6 %	128K	Gemma

Sur le **CyberSecEval benchmark** de Meta — qui évalue la résistance aux abus cybersécurité et la capacité d'assistance défensive — Phi-4 et Gemma 3 27B obtiennent les meilleurs scores de refus appropriés (94 % et 92 % respectivement). L'étude de ces benchmarks spécialisés est cruciale avant tout déploiement dans un contexte SOC, car un modèle trop permissif peut assister des acteurs malveillants, tandis qu'un modèle trop restrictif perd son utilité pour les analystes légitimes.

Sécurité et hardening des Small Language Models

Le déploiement d'un SLM en production expose l'organisation à des vecteurs d'attaque spécifiques que les équipes de sécurité doivent maîtriser. L'*injection de prompt* (prompt injection) reste la menace principale : un attaquant insère des instructions malveillantes dans les données traitées par le modèle pour détourner son comportement. Ce risque est particulièrement aigu lorsque le SLM analyse du contenu externe non maîtrisé (emails, tickets ITSM, logs applicatifs).

L'OWASP Top 10 for LLM Applications liste les dix risques prioritaires pour les déploiements de modèles de langage, dont l'injection de prompt en position LLM01, l'insécurité des plugins en LLM07 et l'empoisonnement des données d'entraînement en LLM03. Ces références sont devenues le standard minimum pour toute analyse de risque IA en contexte professionnel.

> **Avertissement sécurité :** Les SLM open-weights peuvent être fine-tunés pour supprimer leurs garde-fous d'alignement. Des modèles comme *WizardLM-Uncensored* ou *Dolphin-Mistral* ont délibérément retiré les refus RLHF et circulent librement. Ces modèles ne doivent jamais être déployés dans un environnement d'entreprise sans contrôle d'accès strict et audit des usages.

Les mesures de hardening recommandées pour un déploiement SLM enterprise en 2026 incluent :

- **Sandboxing de l'inférence** : exécution du moteur d'inférence dans un conteneur isolé (gVisor, Kata Containers) avec seccomp et AppArmor activés, sans accès réseau sortant.
- **Couche de filtrage applicatif** : proxy NLP entre l'utilisateur et le modèle qui détecte les patterns d'injection de prompt connus (LLM Guard, Rebuff, Lakera Guard).
- **Journalisation des inférences** : conservation de toutes les paires prompt/réponse pour audit forensique, avec chiffrement au repos et durée de rétention alignée sur la politique RGPD.
- **Contrôle d'accès basé sur les rôles** : RBAC sur l'API d'inférence, avec des systèmes prompts différents selon le niveau d'habilitation de l'utilisateur.
- **Validation de l'intégrité du modèle** : vérification du hash SHA-256 des fichiers de poids avant chargement, comparé aux valeurs publiées sur Hugging Face.

La **supply chain des modèles** constitue une surface d'attaque émergente. Des recherches de 2025-2026 ont démontré la faisabilité d'attaques par empoisonnement de poids (weight poisoning) et de backdoor neuraux — des comportements malveillants dormants activés par un trigger spécifique. L'ANSSI recommande dans ses guides IA de ne déployer que des modèles issus de sources vérifiées et idéalement audités par une tierce partie.

Fine-tuning et spécialisation des SLM pour la cybersécurité

La véritable valeur des SLM open-weights réside dans la capacité à les **spécialiser par fine-tuning** sur des corpus métier propriétaires. Pour la cybersécurité, cela signifie entraîner un modèle sur les playbooks de réponse à incident de l'organisation, les règles SIGMA personnalisées, les TTPs observés sur le périmètre surveillé. Notre guide complet sur le fine-tuning LLM 2026 avec LoRA et QLoRA détaille la méthodologie complète.

Le *LoRA (Low-Rank Adaptation)* permet de fine-tuner Phi-4 ou Gemma 3 12B avec seulement **16 à 64 GB de VRAM** en quelques heures, contre des semaines sur cluster GPU pour un entraînement complet. La technique consiste à injecter de petites matrices adaptatives dans les couches d'attention du transformeur, en laissant les poids originaux gelés. Le modèle résultant conserve ses capacités générales tout en développant une expertise spécialisée.

Cas d'usage concrets pour le fine-tuning cybersécurité en 2026 :

- Classification et triage automatique des alertes SIEM avec apprentissage des faux positifs historiques
- Génération de rapports de vulnérabilité au format maison à partir de résultats de scan bruts
- Assistant de threat hunting qui comprend le contexte réseau spécifique de l'organisation
- Analyse de malware statique avec extraction automatique des IOC dans le format STIX/TAXII interne

Déploiement pratique avec Ollama et LM Studio

La **démocratisation du déploiement local** des SLM est portée par des outils qui masquent la complexité de gestion des runtimes d'inférence. Notre article complet sur LLM local avec Ollama, LM Studio et vLLM couvre l'ensemble de l'écosystème. En synthèse pour 2026 :

Ollama (v0.4+) gère nativement Phi-4, Gemma 3, Mistral et Llama 3.x avec une API REST compatible OpenAI. Son système de layers GGUF permet la mutualisation de poids entre

modèles partageant une architecture de base, réduisant l'espace disque nécessaire de 30 à 50 %. La version enterprise introduit un **RBAC natif** et des logs d'audit — fonctionnalités critiques pour les déploiements SOC.

vLLM (v0.6+) s'impose pour les déploiements haute performance avec plusieurs utilisateurs concurrents. Son moteur de *PagedAttention* multiplie par 3 à 5 le débit d'inférence en optimisant la gestion du KV cache. Sur un serveur H100 80GB, vLLM peut servir Gemma 3 27B à 1200 tokens/seconde — suffisant pour alimenter une dizaine d'analystes SOC en parallèle.

Pour les déploiements où la comparaison de modèles est nécessaire, l'approche comparatif LLM open-source 2026 permet d'établir une grille d'évaluation objective avant de valider un modèle pour la production.

À RETENIR

À retenir

- Les **small language models 2026** atteignent des performances de niveau GPT-3.5 à GPT-4 sur des tâches spécialisées, avec une fraction du coût d'inférence.
- **Phi-4** (MIT License) excelle en raisonnement et code ; **Gemma 3** apporte multimodalité et contexte 128K en open-weights.
- L'exécution locale via **GGUF/ONNX** sur NPU Intel/AMD permet des déploiements edge souverains sans dépendance cloud.
- Le fine-tuning **LoRA/QLoRA** permet la spécialisation cybersécurité en quelques heures sur GPU grand public.
- Les risques d'**injection de prompt** et de supply chain IA nécessitent une couche de sécurité applicative dédiée (OWASP LLM Top 10).
- La quantification **Q4_K_M** offre le meilleur compromis performance/taille pour la plupart des déploiements edge.

Architecture RAG avec SLM pour la cybersécurité

Le pattern *RAG (Retrieval-Augmented Generation)* transforme un SLM généraliste en assistant expert capable de répondre sur des bases de connaissances propriétaires sans fine-tuning. Pour une équipe SOC, l'architecture RAG typique combine un SLM local (Phi-4 ou Gemma 3 12B), une base vectorielle (Qdrant, Milvus, pgvector) indexant les playbooks et runbooks, et un pipeline d'ingestion des nouvelles alertes.

L'avantage du RAG sur le fine-tuning est la **mise à jour en temps réel** : les nouvelles TTPs, les nouvelles CVE publiées, les nouveaux IOC peuvent être indexés immédiatement sans réentraînement. Une implémentation RAG typique sur Gemma 3 12B peut répondre à une requête d'analyse en 2-4 secondes avec un contexte de 50 documents retrievés — performances largement suffisantes pour l'assistance à l'analyste.

La combinaison SLM + RAG + quantification GGUF est analysée en profondeur dans notre dossier sur la quantization LLM 2026, notamment les compromis entre précision de la quantification et qualité des embeddings pour la recherche sémantique.

Conformité RGPD, NIS 2 et souveraineté IA

La question de la conformité réglementaire est centrale dans le choix d'un SLM pour les entreprises européennes. L'**EU AI Act**, entré en application en 2026, classe les systèmes d'IA déployés dans des contextes de sécurité critique en risque élevé, imposant transparence, documentation et supervision humaine. Les SLM déployés localement facilitent naturellement cette conformité car les données ne quittent pas le périmètre de l'entreprise.

Le **RGPD** impose une analyse d'impact (DPIA) pour tout traitement de données personnelles par un système IA. Un SLM qui analyse des emails ou des tickets ITSM contenant des données personnelles est soumis à cette obligation. L'exécution locale élimine le risque de transfert hors UE, mais la journalisation des prompts crée elle-même une base de données personnelles à protéger.

La directive **NIS 2**, transposée en droit français, impose aux opérateurs essentiels et importants de maîtriser les risques liés aux technologies qu'ils opèrent, y compris les systèmes IA. Les SLM open-weights auditables répondent mieux à cette exigence que les API cloud propriétaires, dont le comportement interne reste opaque. L'ANSSI a publié en 2025 un guide de sécurisation des systèmes IA qui recommande explicitement l'évaluation des modèles open-source avant déploiement.

FAQ — Small Language Models 2026

Quels sont les avantages des small language models en 2026 par rapport aux grands LLM ?

Les **small language models en 2026** offrent quatre avantages décisifs pour les entreprises. Premièrement, le **coût d'inférence** est réduit de 85 à 95 % : exécuter Phi-4 en local coûte moins d'un centime pour mille tokens, contre 10 à 30 centimes pour GPT-4 Turbo via API. Deuxièmement, la **latence** est incomparablement plus faible sur des requêtes répétitives : 50 ms en local contre 500-2000 ms via une API distante. Troisièmement, la **souveraineté des données** est totale — aucune donnée sensible ne transite par un serveur tiers, ce qui simplifie radicalement la conformité RGPD et NIS 2. Quatrièmement, la **personnalisation par fine-tuning** est accessible et rapide : en 4 à 8 heures sur un GPU RTX 4090, un SLM peut être spécialisé sur les corpus métier de l'organisation, une opération impossible à coût raisonnable sur les LLM propriétaires. Le seul domaine où les SLM restent en retrait est la gestion de tâches complexes et multi-étapes requérant un raisonnement général très étendu, mais ce gap se réduit chaque trimestre.

Comment déployer Phi-4 en environnement Edge sécurisé en 2026 ?

Le déploiement de **Phi-4 en environnement Edge sécurisé** suit une procédure en cinq étapes. Étape 1 — **validation de l'intégrité** : télécharger les poids depuis Hugging Face (microsoft/Phi-4) et vérifier le SHA-256 de chaque fichier contre les valeurs publiées dans le model card. Étape 2 — **quantification** : convertir les poids en GGUF Q4_K_M via llama.cpp pour réduire

l'empreinte mémoire de 28 GB à environ 9 GB. Étape 3 — **containerisation sécurisée** : encapsuler Ollama ou llama.cpp dans un conteneur Podman rootless avec capabilities minimales, seccomp profile strict et réseau désactivé. Étape 4 — **proxy applicatif** : déployer LiteLLM ou un reverse proxy custom pour centraliser l'authentification, la journalisation et le filtrage des prompts. Étape 5 — **monitoring** : configurer des alertes sur les patterns d'injection de prompt, les volumes d'inférence anormaux et les latences dégradées. Cette architecture permet d'opérer Phi-4 sur un mini-serveur edge Intel Core Ultra avec un niveau de sécurité opérationnel satisfaisant pour la plupart des contextes enterprise.

Pourquoi les small language models sont-ils particulièrement pertinents pour la cybersécurité en 2026 ?

La cybersécurité présente un profil d'usage **idéalement adapté aux SLM** pour plusieurs raisons structurelles. Les tâches SOC typiques — classification d'alertes, extraction d'IOC, corrélation de logs, rédaction de rapports — sont des tâches bornées qui n'exigent pas la connaissance encyclopédique d'un GPT-4. Un SLM spécialisé sur la taxonomie MITRE ATT&CK et les playbooks internes surpassera un LLM généraliste sur ces tâches précises. Les **contraintes de confidentialité** sont également déterminantes : les logs de sécurité contiennent des informations sur l'architecture interne, les vulnérabilités et les incidents en cours — les envoyer vers une API cloud constitue un risque inacceptable pour la plupart des RSSI. Enfin, la cybersécurité opérationnelle exige une **disponibilité offline** : en cas d'incident majeur impactant la connectivité, un SLM local continue de fonctionner là où une dépendance API deviendrait un point de défaillance supplémentaire. Ces trois facteurs — précision spécialisable, confidentialité, résilience — font des SLM la technologie IA la plus pertinente pour les équipes de défense en 2026.

Conclusion

Les **small language models 2026** marquent la maturité d'un paradigme : l'intelligence artificielle performante n'est plus l'apanage des hyperscalers. Phi-4 avec sa licence MIT et ses performances de raisonnement exceptionnelles, Gemma 3 avec sa multimodalité et son contexte de 128K

tokens, Mistral avec son positionnement souverain européen — ces modèles offrent dès aujourd'hui des capacités qui auraient nécessité des ressources de datacenter il y a deux ans. Pour les équipes cybersécurité, la combinaison SLM + fine-tuning LoRA + déploiement Edge représente une transformation profonde de l'outillage défensif : des assistants d'analyse locaux, souverains, spécialisés et économiques. La prochaine étape — les SLM multimodaux capables d'analyser captures réseau, screenshots de malware et logs visuels — est déjà engagée avec Gemma 3 et les modèles vision-language compacts. L'enjeu pour 2027 sera la généralisation des agents SLM autonomes capables d'orchestrer des workflows de réponse à incident complets.