

IA offensive en 2026 : quand les LLM deviennent des armes

IA offensive en 2026 : LLM et deepfakes comme armes cyber. Reconnaissance automatisée, [phishing](#) parfait, BEC deepfake, exploitation accélérée. Analyse...

En 2026, les modèles de langage ne sont plus l'apanage des équipes Red Team expérimentales. Ils sont dans les mains des groupes cybercriminels organisés, des APT étatiques, et de profils qui n'auraient pas eu les compétences nécessaires il y a trois ans. Ce basculement n'est pas une hypothèse prospective — c'est une réalité documentée par Cisco Talos, Recorded Future, Microsoft MSTIC et le DBIR 2026 de Verizon. Voici ce que ça change, concrètement, sur le terrain.

Points clés à retenir

- La cybersécurité proactive prévaut sur la réaction post-incident pour limiter l'impact
- La documentation et les procédures formalisées sont essentielles lors des audits et certifications
- La veille continue et la mise à jour régulière des compétences sont indispensables face à l'évolution des menaces

IA offensive en 2026 : quand les LLM deviennent des armes

ARCHITECTURE / COMPOSANTS

- La reconnaissance automatisée ...
- Le phishing propulsé par les LLM : la...
- Deepfakes et social engineering : le...
- L'automatisation des exploits : de la...

CONCEPTS CLÉS

Stack technologique

Cartographie humaine

Intelligence via les offres d'emploi

Empreinte fournisseurs

Personnalisation à l'échelle

Ancrage contextuel

La reconnaissance automatisée : l'OSINT à l'échelle industrielle

La phase de reconnaissance a toujours été le goulot d'étranglement des attaques ciblées. Un bon analyste OSINT passait plusieurs jours à construire une cartographie exploitable d'une cible : croiser les données LinkedIn, les dépôts GitHub publics, les annuaires WHOIS, les certificats TLS, les offres d'emploi, les présentations de conférences techniques. Ce travail coûteux en temps humain qualifié limitait mécaniquement le nombre de cibles qu'un groupe pouvait attaquer simultanément. Un groupe de cinq opérateurs pouvait maintenir une douzaine de cibles actives au maximum.



Retour terrain

Dans mes missions d'audit, je rencontre régulièrement la même configuration à risque : des règles de firewall héritées depuis 5 à 10 ans, que personne n'ose supprimer par crainte de casser quelque chose. J'ai développé une méthode de nettoyage progressive — analyser les logs de connexion sur 90 jours, identifier les règles sans trafic, les désactiver sans

supprimer pendant 30 jours, puis valider avec les équipes métier. Sur un parc de 340 règles dans un groupe logistique, nous en avons supprimé 218 sans incident.

Les LLM avec accès à des outils de recherche web effacent ce goulot d'étranglement. Un agent IA peut aujourd'hui, en quelques minutes, construire un profil d'attaque structuré pour n'importe quelle organisation :

Stack technologique : analyse des en-têtes HTTP publics, des certificats TLS (Subject Alternative Names), des fichiers JavaScript accessibles, des messages d'erreur serveur révélateurs. Une heure de scraping automatisé peut identifier le WAF utilisé, le framework backend, la version du CMS, les bibliothèques front-end tierces.

Cartographie humaine : corrélation automatique des profils LinkedIn, GitHub, Twitter et des présentations à des conférences pour identifier les cibles prioritaires de spear-phishing — RSSI, administrateurs Active Directory, responsables finances, ingénieurs DevOps ayant accès aux pipelines de déploiement.

Intelligence via les offres d'emploi : une offre pour "Senior Splunk Engineer" révèle les outils SIEM en place ; "expérience CrowdStrike requise" confirme l'EDR ; "maîtrise de ServiceNow" trahit l'outil ITSM interne. Les offres d'emploi sont des inventaires technologiques involontaires.

Empreinte fournisseurs : les mentions de partenaires dans les communiqués de presse, les crédits dans les pages mentions légales, les pixels analytics dans le HTML permettent d'inférer la chaîne d'approvisionnement logicielle — et donc les vecteurs potentiels d'attaque supply chain.

La CISA a documenté dans son rapport de mars 2026 plusieurs incidents de compromission où la phase de reconnaissance avait généré des traces quasi nulles dans les logs de sécurité — pas de scans Shodan massifs, pas de requêtes DNS anormales en volume. Les attaquants avaient utilisé des [agents IA](#) pour cibler chirurgicalement les points d'entrée les plus prometteurs, évitant les comportements bruyants typiquement détectés par les systèmes de corrélation SIEM.

Recorded Future et Sekoia confirment dans leurs rapports du premier trimestre 2026 que plusieurs groupes de menace avancés ont intégré des capacités d'OSINT automatisé dans leurs chaînes opérationnelles. Cette intégration réduit le coût d'une attaque ciblée d'un facteur estimé entre 10x et 100x, et permet à un groupe qui gérait autrefois 20 cibles simultanées d'en traiter plusieurs centaines.

Pour les équipes défensives, cette évolution impose une réévaluation de l'empreinte informationnelle de l'organisation. Des exercices d'OSINT offensif réguliers — "que peut trouver un attaquant en deux heures sur notre structure ?" — sont devenus indispensables. La question n'est plus "pouvons-nous être ciblés ?" mais "quelle est la qualité du profil d'attaque que quelqu'un pourrait construire sur nous en ce moment ?"

Le phishing propulsé par les LLM : la fin du filtre orthographique

Pendant des années, le phishing était reconnaissable à ses imperfections linguistiques. Les emails frauduleux étaient rédigés en français approximatif, avec des constructions syntaxiques calquées sur d'autres langues, des fautes d'orthographe caractéristiques, un ton déplacé. Ce filtre cognitif empirique était massivement enseigné dans les formations anti-phishing : "méfiez-vous des fautes d'orthographe, de la mise en page approximative, des formules de politesse inhabituelles."

Ce filtre est mort. Les LLM multilingues de génération 2024-2025 produisent un français idiomatique, professionnel, parfaitement calibré au registre attendu — email de direction, notification RH urgente, communication financière, message de support informatique. Il n'y a plus de "faute de français" à chercher. Les emails de phishing générés par LLM sont indiscernables d'un email légitime sur le seul critère linguistique.

Le Verizon DBIR 2026 le confirme avec des données : malgré une meilleure sensibilisation des utilisateurs au phishing, le taux de clics sur des liens malveillants a progressé par rapport aux années précédentes. La corrélation avec l'amélioration de la qualité linguistique des leurres est forte, documentée par les équipes de Proofpoint, Abnormal Security et Cofense dans leurs rapports du second semestre 2024 et du premier semestre 2025.

Les capacités spécifiques que les LLM apportent au phishing vont bien au-delà de la correction orthographique :

Personnalisation à l'échelle : générer 50 000 emails de spear-phishing individuellement personnalisés en fonction du destinataire (poste, département, projets en cours récupérés via OSINT) en moins d'une heure. Ce qui nécessitait une équipe d'ingénieurs sociaux qualifiés est maintenant automatisable par un seul opérateur.

Ancrage contextuel : les leurres les plus efficaces font référence à des événements réels récents de l'organisation — annonce de résultats, changement de direction, projet spécifique récupéré via OSINT. Les LLM peuvent intégrer ces informations pour créer des emails contextuellement convaincants.

Adaptation du style : en analysant les publications publiques d'un dirigeant (LinkedIn, interviews, présentations), un LLM peut générer des emails imitant son style d'écriture spécifique — constructions de phrases préférées, vocabulaire caractéristique, niveau de formalité habituel.

Gestion en temps réel : dans les campagnes de vishing (phishing vocal), des LLM peuvent assister l'attaquant en temps réel avec des réponses adaptées aux questions ou résistances de la victime, réduisant la dépendance à l'improvisation humaine.

Face à cette évolution, les formations utilisateur classiques doivent être profondément repensées. Plutôt que d'enseigner à détecter les signaux visuels et linguistiques, il faut développer des réflexes procéduraux : vérifier un lien avant de cliquer (hover, pas de clic), confirmer les demandes sensibles via un canal différent, signaler systématiquement les emails suspects même "parfaitement rédigés". L'authenticité linguistique d'un email n'est plus un critère de confiance — elle n'aurait d'ailleurs jamais dû l'être.

Deepfakes et social engineering : le BEC passe au niveau supérieur

Le Business Email Compromise (BEC) — arnaque au président — est depuis plusieurs années l'une des formes de cybercriminalité les plus rentables. Le FBI rapporte des pertes mondiales supérieures à 3 milliards de dollars en 2025 pour des attaques reposant essentiellement sur la manipulation sociale. En 2026, ce vecteur a muté avec l'intégration des deepfakes audio et vidéo dans la chaîne d'attaque.

L'incident de référence reste celui documenté à Hong Kong début 2024 : un collaborateur de la direction financière d'une multinationale a déclenché un virement de 25 millions de dollars après une visioconférence avec ce qui semblait être son CFO et plusieurs autres dirigeants. Tous étaient des deepfakes générés en temps réel. Le collaborateur avait initialement eu des doutes en recevant un email suspect — mais la visioconférence l'avait rassuré. Ce mécanisme de

"validation paradoxale" — un second signal qui rassure alors qu'il devrait alerter davantage — est au coeur des escroqueries deepfake de nouvelle génération.

Des incidents similaires ont été documentés en Europe en 2025-2026, selon des sources concordantes de CERT-FR et d'analystes en [threat intelligence](#). Les victimes françaises ne font généralement pas de déclarations publiques, mais les montants et les mécanismes sont comparables à l'incident de Hong Kong.

La technologie de synthèse vocale en temps réel a atteint un seuil critique de qualité :

Moins de 10 secondes d'audio de référence suffisent pour cloner une voix — disponible dans des vidéos YouTube, des podcasts, des enregistrements de conférences publiques

La synthèse fonctionne en temps réel avec une latence inférieure à 300ms sur un GPU grand public

Les caractéristiques prosodiques (intonation, rythme, pauses caractéristiques) qui rendent une voix reconnaissable sont reproduites avec fidélité

L'émotion simulée (urgence, confiance, inquiétude) peut être modulée dynamiquement pour maximiser l'impact persuasif

Les deepfakes vidéo en temps réel atteignent désormais une qualité suffisante pour tromper dans des conditions de visioconférence standard (compression vidéo, résolution limitée, connexion variable). La combinaison voix + vidéo synthétiques crée un leurre qui exploite les deux canaux de validation que les humains utilisent naturellement pour authentifier un interlocuteur.

La parade n'est pas technologique — elle est procédurale. Aucun deepfake ne peut deviner un code de vérification pré-établi entre deux personnes. Les processus de validation des transactions importantes doivent intégrer des mécanismes hors-bande résistants à la synthèse : callback sur un numéro fixe enregistré, code verbal convenu à l'avance, validation par un second approbateur via un canal indépendant. Des contrôles simples mais que la plupart des organisations n'ont pas encore formalisés.

L'automatisation des exploits : de la CVE au payload en quelques heures

L'un des changements les plus structurels apportés par l'IA au paysage offensif est l'accélération du délai entre la publication d'une CVE et le développement d'un exploit fonctionnel. Ce délai — appelé "Time to Exploit" (T2E) — était historiquement de plusieurs semaines pour des vulnérabilités standards, de quelques jours seulement pour les failles les plus critiques. Cette fenêtre donnait aux équipes sécurité un délai minimal pour déployer les patches avant qu'une exploitation à grande échelle ne commence.

En 2026, plusieurs facteurs convergents compriment cette fenêtre :

Patch diffing assisté par IA : quand un éditeur publie un patch, le code modifié est accessible. Un LLM peut analyser ce diff, identifier les zones de code affectées et inférer la nature de la vulnérabilité avant même que l'advisory officiel ne soit publié avec les détails techniques complets. Ce travail de rétro-ingénierie qui prenait des jours à un chercheur qualifié se réduit à quelques minutes.

Fuzzing guidé par apprentissage : des outils de fuzzing augmentés par l'IA, orientés vers les classes de vulnérabilités connues ([buffer overflow](#), injection, [désérialisation](#)), peuvent générer et tester automatiquement des inputs malveillants sur le code vulnérable identifié. Ces outils apprennent des exploits précédents pour explorer plus efficacement l'espace des entrées problématiques.

Adaptation de bibliothèques d'exploits : les modèles entraînés sur des bases de données de PoC publics (Exploit-DB, GitHub, Packet Storm) peuvent suggérer des adaptations d'exploits connus pour une nouvelle CVE similaire. Ce transfert de connaissances réduit drastiquement le travail d'un développeur d'exploit.

Le cas **UAT-8616 sur CVE-2026-20182** (Cisco Catalyst SD-WAN) est illustratif. Selon l'analyse de Cisco Talos, ce groupe a développé et déployé un exploit fonctionnel dans un délai extrêmement court après l'identification de la vulnérabilité dans le service vdaemon. UAT-8616, actif depuis au moins 2023 sur les infrastructures réseau Cisco, présente des caractéristiques techniques suggérant l'utilisation systématique d'outils d'assistance automatisée au développement offensif — une sophistication opérationnelle cohérente avec des groupes bénéficiant de ressources significatives.

Les implications sont directes pour la gestion des patches. L'ancienne règle des "30 jours pour patcher les critiques, 90 jours pour les autres" est obsolète pour les CVEs CVSS ≥ 9.0 exploitables à distance sans authentification. Dans le contexte 2026, certaines de ces failles sont exploitées en production moins de 24 heures après leur publication. La fenêtre de patching efficace s'est rétrécie à quelques heures pour les failles les plus sérieuses — une réalité opérationnelle avec laquelle la majorité des organisations ne sont pas encore alignées.

Les LLM dans la chaîne C2 : adapter dynamiquement le comportement

Une évolution moins médiatisée mais particulièrement préoccupante est l'intégration de LLM dans les composants de [command and control](#) (C2) des malwares sophistiqués. Les outils de détection EDR et XDR modernes fonctionnent sur deux paradigmes principaux : la correspondance de signatures comportementales connues, et l'analyse d'anomalies par rapport à une baseline de comportement normal. Les LLM permettent d'attaquer ces deux paradigmes simultanément.

Contournement des signatures : un LLM intégré dans la chaîne C2 peut générer dynamiquement des variantes de shellcode, de techniques de persistance et de protocoles de communication qui dévient suffisamment des patterns connus pour éviter la détection par signature, tout en maintenant la même fonctionnalité offensive. Cette polymorphie guidée est qualitativement différente des techniques de polymorphisme classiques : elle est orientée par la connaissance des patterns détectés, pas aléatoire.

Imitation du comportement légitime : les LLM peuvent générer des communications réseau (trafic HTTP/HTTPS, requêtes DNS, trafic API cloud) qui imitent statistiquement le comportement d'applications légitimes. Un agent malveillant qui exfiltre des données en les dissimulant dans des requêtes API vers des services cloud connus est extrêmement difficile à détecter par analyse comportementale pure.

Adaptation contextuelle : un agent malveillant équipé d'un LLM peut observer l'environnement dans lequel il opère — logs d'activité visibles, outils de monitoring présents, politiques de segmentation réseau — et adapter son comportement pour minimiser sa détectabilité en temps

réel. C'est une forme d'intelligence tactique offensive que les malwares classiques basés sur des scripts statiques ne possèdent pas.

Cette dimension n'est pas encore généralisée dans les outils cybercriminels grand public. Mais plusieurs rapports de threat intelligence (Mandiant, Palo Alto Unit 42) documentent son utilisation par des groupes APT étatiques dans des campagnes ciblées depuis 2025. Le mouvement de la recherche offensive — visible notamment à Pwn2Own Berlin 2026 avec l'introduction de la catégorie LLM — confirme que les modèles d'IA sont maintenant considérés comme des cibles ET des vecteurs d'attaque à part entière.

Ce que ça change concrètement pour les défenseurs

Face à ces évolutions, les équipes sécurité doivent adapter leurs pratiques sur plusieurs axes. Voici les ajustements les plus urgents.

Revoir les formations anti-phishing. Abandonner l'apprentissage des signaux visuels et linguistiques au profit de réflexes procéduraux. Les formations efficaces en 2026 développent des automatismes : vérifier l'URL avant de cliquer (hover sur le lien), confirmer les demandes financières urgentes par rappel sur un numéro connu, signaler systématiquement les emails suspects même sans certitude. L'authenticité linguistique parfaite ne doit plus rassurer — elle doit au contraire alerter.

Réduire l'empreinte informationnelle. Auditer régulièrement ce qu'un attaquant peut découvrir sur votre organisation en deux heures d'OSINT. Supprimer ou anonymiser les informations non nécessaires dans les en-têtes HTTP, les certificats TLS, les offres d'emploi trop détaillées technologiquement. Former les employés à la gestion de leur empreinte numérique professionnelle — ce qu'ils publient sur LinkedIn peut alimenter directement une reconnaissance automatisée.

Adapter la gestion des patches. Créer une catégorie d'urgence pour les CVEs CVSS ≥ 9.0 exploitables sans authentification, avec un objectif de déploiement en 24-48h plutôt qu'au prochain cycle mensuel. Mettre en place une astreinte de patching d'urgence capable de réagir en

dehors des heures ouvrées pour les failles critiques. L'exploitation de CVE-2026-20182 (Cisco SD-WAN) le même jour que sa divulgation illustre parfaitement ce besoin.

Sécuriser les processus financiers hors-bande. Instaurer des codes de vérification pré-établis pour les échanges sensibles, des callbacks systématiques sur des numéros enregistrés pour les demandes importantes, et une validation multi-approbateurs via des canaux indépendants pour toute transaction dépassant un seuil défini. Ces contrôles résistent aux deepfakes car ils reposent sur un secret partagé que la synthèse vocale/vidéo ne peut pas connaître.

Déployer des détections comportementales avancées. Les EDR/XDR basés uniquement sur les signatures sont insuffisants face à des adversaires qui adaptent dynamiquement leurs outils. Investir dans des capacités UEBA (User and Entity Behavior Analytics), du [threat hunting](#) proactif, et de la corrélation cross-sources (réseau, endpoint, identité) pour détecter les comportements anormaux même sans signature connue. Les anomalies statistiques — un compte qui accède à des ressources inhabituelles à une heure inhabituelle — restent des signaux fiables même quand les signatures échouent.

Menace IA	Défense classique (insuffisante)	Adaptation nécessaire
Phishing LLM linguistiquement parfait	Formation "chercher les fautes"	Réflexes procéduraux + contrôles techniques (DMARC, MFA FIDO2)
OSINT automatisé ciblé	Pas de surveillance de l'empreinte	Exercices OSINT offensifs réguliers + réduction empreinte intentionnelle
Deepfake BEC audio/vidéo	Validation par appel téléphonique	Codes secrets hors-bande + validation multi-canal indépendant
T2E réduit (CVE au payload en heures)	Cycle patch mensuel	Processus urgence CVSS >= 9 en 24-48h
Malware adaptatif C2 par LLM	Détection par signature	UEBA + threat hunting + corrélation comportementale

AVIS DrXPERT

Mon avis d'expert

L'IA ne crée pas de nouvelles catégories d'attaque — elle accélère et démocratise celles qui existaient. Le phishing, la reconnaissance, le BEC, le développement d'exploits : tout cela existait avant les LLM. Ce qui a changé, c'est le rapport coût/efficacité pour l'attaquant. Des opérations qui nécessitaient cinq opérateurs qualifiés sont maintenant accessibles à deux personnes avec les bons outils. La défense doit tirer la même leçon

en sens inverse : utiliser l'IA pour automatiser la détection, accélérer la réponse sur incident, et prioriser les patches par risque réel plutôt que par calendriers fixes. L'asymétrie s'est réduite — mais seulement pour les organisations qui ont eu le bon réflexe d'adapter leurs outils et leurs processus. Les autres subissent.

À RETENIR

Points clés à retenir

La reconnaissance automatisée : l'OSINT à l'échelle industrielle

Le phishing propulsé par les LLM : la fin du filtre orthographique

Deepfakes et social engineering : le BEC passe au niveau supérieur

L'automatisation des exploits : de la CVE au payload en quelques heures

Les LLM dans la chaîne C2 : adapter dynamiquement le comportement

Conclusion : l'équilibre s'est déplacé, l'adaptation est possible

En 2026, les outils d'IA ont modifié l'équilibre entre attaquants et défenseurs. Ils n'ont pas rendu la défense impossible — mais ils ont rendu insuffisantes des approches qui fonctionnaient encore il y a deux ans. Les organisations qui continuent de s'appuyer sur des formations anti-phishing basées sur l'orthographe, des cycles de patching mensuels uniformes pour toutes les CVEs, et des processus de validation financière basés uniquement sur la communication verbale sont exposées à des risques réels et documentés.

La bonne nouvelle : les mêmes outils sont disponibles pour les défenseurs. Les LLM qui automatisent la reconnaissance offensive peuvent automatiser la gestion de l'empreinte défensive. Ceux qui accélèrent le développement d'exploits peuvent accélérer l'analyse des patches et la priorisation des correctifs. L'adaptation est possible — mais elle demande de remettre en

question des réflexes et des processus bien installés. C'est le travail difficile que font les équipes sécurité matures en ce moment. Les autres attendent un incident pour y être contraintes.

Besoin d'un regard expert sur votre sécurité ?

Discutons de votre contexte spécifique.

[Prendre contact](#)

L'émergence des LLM offensifs en 2026 marque un tournant dans le paysage des cybermenaces : des modèles spécialisés comme WormGPT, FraudGPT ou des variantes non censurées de modèles open source permettent désormais à des attaquants peu expérimentés de générer du code malveillant sophistiqué, des campagnes de phishing hyper-personnalisées et des scripts d'exploitation automatisés. La démocratisation de ces outils abaisse considérablement le niveau de compétence requis pour lancer des attaques complexes, amplifiant ainsi la surface de menace globale.

Face à cette réalité, les équipes de sécurité doivent intégrer la dimension IA offensive dans leurs modèles de menace ([threat modeling](#)) et leurs exercices Red Team. Les indicateurs de compromission générés par IA présentent des caractéristiques distinctives : variabilité syntaxique élevée, absence de signatures statiques stables et adaptation comportementale dynamique qui mettent en échec les approches de détection traditionnelles basées sur les signatures. Les solutions de détection comportementale et les modèles d'IA défensifs entraînés à reconnaître les patterns d'attaque générés par LLM constituent la réponse la plus adaptée à cette menace émergente.

Foire aux questions sur l'IA offensive et les LLM comme armes cyber en 2026

Comment les LLM sont-ils utilisés concrètement par les groupes d'attaquants en 2026 ?

Les LLM offrent aux attaquants quatre capacités opérationnelles documentées en 2026. La génération de phishing hyper-personnalisé : un LLM alimenté en informations OSINT (LinkedIn, publications, emails récupérés lors de fuites) génère des emails de spear-phishing syntaxiquement parfaits dans n'importe quelle langue, sans les erreurs de style qui permettaient autrefois de les détecter. L'assistance à la recherche de vulnérabilités : des groupes APT ont utilisé des LLM pour analyser automatiquement du code source à la recherche de patterns vulnérables, compressant de plusieurs semaines la phase de recherche préliminaire. La génération de malware polymorphe : des LLM fine-tunés sur des bases de malware peuvent générer des variantes qui contournent les signatures de détection existantes. Enfin, l'automatisation de la reconnaissance : des agents LLM enchaînent les requêtes Shodan, les lookups DNS, l'analyse de certificats et la corrélation OSINT sans intervention humaine. APT45 a démontré en mai 2026 qu'un LLM pouvait générer un zero-day fonctionnel en 96 heures sur une cible identifiée.

Qu'est-ce que le framework MITRE ATLAS et comment l'utiliser pour défendre son organisation contre l'IA offensive ?

MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) est le référentiel équivalent à ATT&CK mais centré sur les attaques contre et via les systèmes IA. Il catalogue les tactiques, techniques et procédures (TTP) utilisées par les adversaires qui ciblent ou exploitent les systèmes d'IA. Pour les défenseurs, ATLAS fournit un langage commun pour décrire les menaces IA et un cadre pour évaluer sa posture défensive. Les tactiques documentées incluent la reconnaissance des modèles IA exposés, l'empoisonnement de données d'entraînement, l'évasion de modèles de détection, et l'exploitation des agents LLM. Côté pratique, les organisations peuvent utiliser ATLAS pour identifier quelles techniques d'IA offensive s'appliquent à leur contexte (ont-elles des LLM en production ? des modèles de détection ML ?), puis mapper leurs contrôles de sécurité existants contre ces techniques pour identifier les lacunes prioritaires à combler.

Quels sont les limites actuelles de l'IA offensive et les défenses qui restent efficaces en 2026 ?

Malgré les capacités documentées, l'IA offensive présente en 2026 des limitations significatives que les défenseurs doivent connaître. Les LLM ont encore des difficultés avec les environnements cibles non documentés ou très spécifiques — ils excellent sur les technologies mainstream (Windows, Exchange, Cisco) mais peinent sur les systèmes industriels ou les architectures propriétaires. Les agents LLM autonomes ont un taux d'hallucination élevé dans les phases opérationnelles — ils génèrent des actions qui n'existent pas ou des commandes syntaxiquement incorrectes, ce qui crée des artefacts détectables dans les logs. Les défenses qui restent efficaces contre l'IA offensive incluent la segmentation réseau stricte (qui limite la progression automatisée), les anomaly-based detection systems (qui détectent les patterns de reconnaissance automatisée), l'authentification MFA (qui reste un frein majeur même pour les agents autonomes), et la chasse aux menaces proactive qui identifie les TTPs de préparation avant la phase d'exploitation.