SPOTLAB / ANALYSE P09

Sécurité de l'IA générative en 2026 : ce que révèle OWASP et comment agir

Ce que change l'OWASP 2026 et comment sécuriser les données, outils et agents grâce à cinq frontières de confiance vérifiables.



Analyse / P09

Novembre 2026

Sources consultées au 2 novembre 2026

ANALYSE P09

Sommaire

01	Trois enseignements pour les décideurs	3
02	Ce qui change dans l'OWASP Top 10 LLM 2026	4
03	Pourquoi les agents exigent une lecture supplémentaire	5
04	Une méthode Spotwork.AI : sécuriser cinq frontières de confiance	6
05	Cas concret : une source documentaire malveillante	8
06	Comment décider qu'un usage peut passer en production ?	9
07	Gouvernance : des preuves techniques aux décisions métier	10
08	Ce que cette analyse ne démontre pas	10
09	Conclusion : sécuriser l'autonomie, pas seulement la réponse	11
10	Sources principales	12

FICHE DE L'ANALYSE

ÉDITEUR
Spotwork.AI — Spotlab, Research & Insights

RÉFÉRENCE
P09

TYPE
Analyse documentaire et cadre méthodologique

COLLECTION
Fiabilité & gouvernance

SOURCES CONSULTÉES
2 novembre 2026

PUBLICATIONS LIÉES
I0-02 · P08

EN BREF

La sécurité d'une application d'IA ne se réduit ni au choix d'un modèle fiable, ni à un « prompt de sécurité ». Dès qu'un assistant lit des documents, récupère des connaissances ou agit sur des systèmes, il traverse des frontières de confiance. La version 2026 du Top 10 OWASP pour les applications LLM et son référentiel consacré aux agents invitent à traiter ensemble les données, les droits, les outils et les traces d'exécution. Pour un dirigeant, l'enjeu est simple : quelles actions le système peut-il réellement accomplir — y compris lorsqu'une source tente de le manipuler ?

> *La sécurité ne se limite pas au modèle : accès, sources, outils, données et actions doivent être examinés ensemble.*
>
> CE QU'IL FAUT RETENIR

01 Trois enseignements pour les décideurs

1

Le risque se situe dans le système complet.

Une réponse pertinente n'exclut pas une fuite de données, une action hors périmètre ou une consommation incontrôlée. Modèle, recherche documentaire, outils, API et identité doivent être évalués ensemble.

2

Les instructions ne remplacent pas les contrôles.

Un document externe peut contenir une tentative d'injection de prompt. La séparation des données et des instructions est utile, mais l'accès aux informations et l'autorisation d'agir doivent être appliqués par des composants qui ne dépendent pas de la bonne volonté du modèle.

3

La sécurité doit être prouvée dans la durée.

Un résultat obtenu en démonstration ne garantit pas la résistance aux changements de modèle, de configuration, de sources ou de permissions. Les tests adversariaux doivent être représentatifs, traçables et régulièrement rejoués.

CE QUI DÉPLACE LE PÉRIMÈTRE DE CYBERSÉCURITÉ

Les informations récupérées ne sont plus seulement consultées : elles peuvent influencer les décisions d'un modèle, et certains modèles peuvent ensuite solliciter des outils qui écrivent, transmettent ou déclenchent des actions. **L'exposition réelle dépend donc moins d'une capacité abstraite du LLM que du couplage entre entrées, contexte, identité, connecteurs et exécution.**

02 Ce qui change dans l'OWASP Top 10 LLM 2026

L'édition 2026 officielle, publiée en août, réorganise les dix familles de risques. Elle place Excessive Agency au troisième rang (contre le sixième en 2025) et introduit Hidden Context Exposure, catégorie plus large qui remplace l'ancienne « System Prompt Leakage » [1][7]. Le classement est un cadre de sensibilisation élaboré par une communauté d'experts, pas une estimation statistique de la probabilité d'une attaque dans votre organisation.

Catégorie officielle 2026	Question opérationnelle à poser	Angle de vérification prioritaire
LLM01 — Prompt Injection	Une donnée externe peut-elle détourner le comportement attendu ?	Séparer contenu non fiable, instructions et décisions d'action
LLM02 — Sensitive Information Disclosure	Le système peut-il divulguer une information interdite ?	Isoler les données et tester la divulgation non autorisée
LLM03 — Excessive Agency	Dispose-t-il de plus d'actions ou de privilèges que nécessaire ?	Réduire la portée des outils, identités et privilèges
LLM04 — Supply Chain	Quels modèles, dépendances, outils et fournisseurs sont maîtrisés ?	Inventorier modèles, connecteurs, bibliothèques et prestataires
LLM05 — Data and Model Poisoning	Une source ou un composant manipulé peut-il corrompre le résultat ?	Contrôler provenance, intégrité et mises à jour des corpus
LLM06 — Unbounded Consumption	Qui limite les appels, ressources, coûts et enchaînements ?	Borner appels, budgets, durée et ressources
LLM07 — Misinformation	Comment les erreurs de contenu sont-elles détectées et traitées ?	Vérifier la véracité et prévoir une supervision métier selon l'impact
LLM08 — Hidden Context Exposure	Quelles informations de contexte pourraient être révélées ?	Limiter et vérifier l'exposition des informations internes au contexte
LLM09 — Vector and Embedding Weaknesses	Le dispositif de recherche vectorielle respecte-t-il les périmètres d'accès ?	Tester recherche, isolation des espaces et filtrage des autorisations
LLM10 — Improper Output Handling	La sortie est-elle contrôlée avant d'être utilisée par un autre système ?	Valider les sorties avant leur exécution ou ingestion en aval

Cette nomenclature est celle du dépôt canonique OWASP 2026 [2]. Elle ne se substitue pas à une analyse de risques tenant compte de l'architecture, des données, des métiers et des impacts. Les angles de vérification sont une traduction pragmatique de Spotwork.AI : ils ne reprennent pas exhaustivement les mesures des fiches OWASP et ne constituent pas un mapping officiel de conformité. Le classement ne doit pas être utilisé comme une probabilité de survenance ou une mesure de gravité universelle : pour chaque catégorie, la criticité se calcule à partir de l'usage, de la capacité du système et des conséquences plausibles.

03 Pourquoi les agents exigent une lecture supplémentaire

Un assistant qui produit uniquement une proposition de texte et un agent autorisé à envoyer des courriels, modifier un fichier ou déclencher un traitement ne présentent pas le même impact potentiel. L'OWASP Top 10 for Agentic Applications 2026 examine spécifiquement la délégation et l'autonomie. Son édition porte le millésime 2026, mais a été annoncée en décembre 2025 [3].

L'Agent Control Standard, accueilli au sein du projet OWASP en septembre 2026, met l'accent sur l'inspection et l'application des règles au moment de l'exécution. La spécification n'en est qu'à sa version 0.1 : elle se prête à des pilotes plutôt qu'à un déploiement généralisé [4][7]. Cela ne signifie pas qu'une entreprise doit adopter un produit ou une architecture précise : le principe est de rendre les actions observables et contrôlables hors du seul dialogue avec le modèle.

Ce que chaque référentiel permet — et ne permet pas — d'affirmer

OWASP Top 10 for LLM Applications 2026

Référentiel communautaire des grandes familles de faiblesses des applications LLM. Les identifiants LLM01 à LLM10 documentent des catégories, pas une probabilité ni une gravité universelle.

OWASP Top 10 for Agentic Applications 2026

Approche centrée sur la délégation, les outils et les systèmes capables de poursuivre des objectifs. Le chevauchement avec Excessive Agency n'autorise pas à fusionner les référentiels ni à additionner artificiellement leurs risques.

Agent Control Standard (ACS)

Principe d'inspection et d'application de politiques au moment de l'exécution, par des points d'ancrage (« hooks ») exposés par les plateformes d'agents. Il ne dispense pas de l'architecture d'identité, des politiques d'accès et des validations métier.

NIST AI 100-2e2025

Taxonomie officielle de l'apprentissage automatique adversarial, qui classe notamment les attaques selon l'atteinte à la disponibilité, à l'intégrité, à la vie privée et, pour l'IA générative, le mésusage [5]. Elle formalise un vocabulaire ; elle ne déclare pas qu'une protection unique élimine tous les risques.

NSA/CISA et partenaires — Deploying AI Systems Securely (avril 2024)

Ce guide conjoint de déploiement sécurisé préconise entre autres des contrôles d'API, des accès stricts, des journaux, de la surveillance et des tests réguliers [6]. Ces mesures de cybersécurité classiques conservent leur rôle dans les systèmes d'IA.

04 Une méthode Spotwork.AI : sécuriser cinq frontières de confiance

Nous proposons une grille opérationnelle inspirée des risques recensés par OWASP et des recommandations de sécurité du NIST et des agences de cybersécurité partenaires de la NSA/CISA. C'est une proposition méthodologique Spotwork.AI, non un standard OWASP et non une certification.

Frontière	Contrôle prioritaire	Preuve attendue
1. Sources ingérées	Marquer l'origine, la fiabilité et le caractère non exécutable des contenus externes	Journal d'ingestion ; tests de contenus adversariaux
2. Données et connaissances	Appliquer les droits d'accès au moment de la recherche et de la restitution	Tests inter-profils et inter-espaces ; décisions d'accès
3. Outils et actions	Déléguer les autorisations à un service déterministe, avec moindre privilège	Traces d'actions autorisées, refusées et approuvées
4. Sorties et destinations	Valider les formats, destinataires et effets avant écriture ou transmission	Journal de validation ; cas négatifs bloqués
5. Changements et supervision	Versionner les configurations, surveiller les anomalies et rejouer les tests	Résultats de régression, incidents, décisions de déploiement

LA RÈGLE STRUCTURANTE

Le modèle peut proposer une opération, mais il ne doit pas être la seule autorité qui décide si celle-ci est permise. Les contrôles dépendent des capacités effectivement exposées : une application sans outil d'action n'a pas besoin du même dispositif qu'un agent capable d'effectuer des transactions.



04 / MODÈLE DE MENACE DÉTAILLÉ PAR FRONTIÈRE

1 Sources ingérées

ACTIFS Documents internes, web, e-mails, échanges et métadonnées.

MENACE Contenu tiers incorporant des instructions déguisées ou une information empoisonnée.

CONTRÔLES Identification des sources, classification de confiance, analyse du contenu, séparation de rôle, impossibilité pour une source d'accorder une nouvelle permission.

PREUVES Enregistrement de la source et de sa version, résultats de tests d'injection directs et indirects, causes de blocage.

2 Données et connaissances

ACTIFS Corpus RAG, index, graphe, pièces jointes, contexte long.

MENACES Restitution inter-tenant, exposition de contexte caché, récupération de documents après révocation d'accès, fuite par traces ou citations.

CONTRÔLES Autorisation au moment de la requête, isolation des espaces, règles documentaires héritées, suppression ou invalidation des copies autorisées, conservation proportionnée des journaux.

PREUVES Cas de test avec comptes et profils distincts, y compris changements de rôle en cours de session.

3 Outils et actions

ACTIFS API métiers, courriels, stockage, systèmes transactionnels.

MENACES Privilèges excessifs, choix d'outil erroné, délégation non autorisée, approbation contournée.

CONTRÔLES Identité explicite de l'utilisateur et du service, politiques « deny by default », capacité minimale, validation des arguments et destinataires, approbation humaine vérifiée pour les actions sensibles, possibilité de révocation et d'arrêt.

PREUVES Journal horodaté des propositions d'action, autorisations, refus et décisions humaines.

4 Sorties et destinations

ACTIFS Documents publiés, messages, fichiers, bases et workflows aval.

MENACES Contenu trompeur, injection vers un autre composant, données non autorisées dans un export.

CONTRÔLES Schémas de données, contrôle des destinations, filtrage des contenus interdits selon le contexte, vérification métier lorsque l'enjeu le nécessite.

PREUVES Essais de formats invalides, de destinataires interdits et d'erreurs de traitement.

5 Changements et supervision

ACTIFS Configuration, modèles, prompts, politiques, index, dépendances.

MENACES Régression silencieuse, changement de fournisseur, hausse de coûts ou journaux incomplets.

CONTRÔLES Versions, tests avant déploiement, détection d'anomalies, limites de consommation, gestion des incidents, restauration et révocation.

PREUVES Résultats horodatés, arbre de décision, responsable de l'exception et audit de remédiation.

05 Cas concret : une source documentaire malveillante

SCÉNARIO ILLUSTRATIF et non incident observé chez Spotwork.AI

Un assistant de cabinet de conseil résume des documents internes avec une page web externe. Cette page contient une instruction qui prétend être prioritaire et demande l'envoi d'un dossier client vers une destination non autorisée.

Un simple test consistant à vérifier si l'assistant répond « je refuse » serait insuffisant. Il faut au minimum tester trois frontières :

FRONTIÈRE 1

La donnée externe n'est pas une instruction privilégiée.

FRONTIÈRE 2

L'accès au dossier client est évalué selon l'identité réelle de l'utilisateur.

FRONTIÈRE 3

L'outil d'envoi refuse toute destination ou action non autorisée.

Une validation humaine peut être exigée pour les transmissions sensibles, mais elle ne doit pas compenser l'absence de contrôle des droits.

Pourquoi le risque est réel

Le modèle peut confondre une donnée consultée avec une instruction opérationnelle. Même si l'interface est conçue pour produire une simple synthèse, un connecteur d'envoi mal protégé peut transformer une génération en divulgation effective.

Dispositif cible

La page récupérée est traitée comme un contenu non fiable. Le service documentaire vérifie que le compte initiateur a le droit de consulter le dossier. Un service distinct valide l'identité, la finalité, le destinataire et les paramètres d'un éventuel envoi. Toute destination non autorisée est refusée quelle que soit l'argumentation produite par le modèle ; une approbation forte reste nécessaire pour certains transferts légitimes.

Test reproductible

Constituer un dossier fictif, deux profils dotés de droits différents, une destination autorisée et une destination non autorisée. Exécuter les cas de base, des documents adversariaux et des variantes. Vérifier non seulement le texte généré, mais les appels aux outils, les décisions d'autorisation, le stockage et les journaux.

> Le résultat attendu n'est pas un « score de sécurité » global : c'est la démonstration que l'action interdite ne peut pas aboutir, y compris si le modèle formule la demande de manière convaincante. Les échecs, hésitations et contournements détectés lors des essais doivent être conservés pour les régressions. Un succès sur le périmètre testé ne constitue pas une garantie absolue.

06 Comment décider qu'un usage peut passer en production ?

Le comité responsable peut demander cinq preuves :

- Inventaire des sources et des outils
- Permissions effectivement appliquées
- Résultats des tests de divulgation et d'injection
- Mécanismes d'approbation et d'arrêt
- Surveillance et plan de réponse à incident

Le périmètre de test doit inclure les cas normaux, les cas négatifs et les scénarios adversariaux plausibles. Le nombre de répétitions ne doit pas être universel : il dépend de la gravité des impacts, de la variabilité du modèle et du coût des essais. En revanche, un seul contournement critique reproductible constitue une raison de bloquer ou de corriger le déploiement concerné, même si les performances moyennes sont bonnes. La décision doit mentionner les limites de couverture des tests.

Protocole de stress test

Périmètre

Décrire précisément le système, les versions de modèle, connecteurs, politiques, index et conditions d'exécution. Garder une référence reproductible et archiver les cas de tests. Toutes les simulations doivent utiliser des données fictives ou autorisées.

Dix familles de tests

1. Injection dans la requête
2. Injection indirecte dans un document
3. Lecture d'un dossier interdit
4. Restitution inter-tenant
5. Modification d'identité ou de rôle
6. Appel d'outil hors périmètre
7. Action sensible sans approbation
8. Sortie malformée envoyée vers un outil aval
9. Consommation excessive
10. Changement de modèle, de règles ou de source après validation

Mesures

Enregistrer la capacité sollicitée, la décision d'autorisation, le résultat réel en aval, la nature de l'échec, la latence et la consommation. Distinguer « le modèle a répondu non » de « l'action interdite a été empêchée par le service ». Mesurer aussi les faux blocages qui rendent le système inutilisable.

Répétition

Fixer avant les essais le nombre de replays et une borne de coût et de temps par scénario, proportionnés au risque et à la variabilité attendue. Un score global moyen ne suffit pas : les refus erronés et les contournements critiques doivent être analysés séparément. L'équipe sécurité définit le seuil d'acceptation ; aucun nombre universel de tests n'est scientifiquement justifié.

Décision de déploiement

GO — PÉRIMÈTRE JUSTIFIÉ | CORRECTION PUIS RETEST | NO-GO — BLOCAGE OU SUSPENSION

Un contournement critique constaté et reproductible bloque le périmètre affecté jusqu'à sa correction et sa nouvelle évaluation. Les cas non testés et les risques résiduels doivent être documentés, avec un responsable décisionnaire.

07 Gouvernance : des preuves techniques aux décisions métier

Acteur	Rôle
RSSI / sécurité produit	Pilote le modèle de menace, l'assurance des permissions et les tests d'intrusion pertinents
Propriétaire métier	Définit les conséquences et les actions qui exigent une validation
Produit / architecture	Rend les contrôles effectifs, journalisables et bornés
Responsable des données	Définit les règles de confidentialité et d'accès
Gouvernance IA	Conserve la trace des décisions et des dérogations ; ne se substitue pas aux contrôles techniques

L'architecture doit pouvoir répondre sans ambiguïté aux questions suivantes :

- ? Quel acteur a demandé quelle action ?
- ? Quelle donnée a déclenché la demande ?
- ? Quelle politique a statué ?
- ? Qui a approuvé, le cas échéant ?
- ? Quelle opération a réellement été réalisée ?
- ? Comment la stopper et la réparer ?

08 Ce que cette analyse ne démontre pas

Le Top 10 OWASP ne garantit ni exhaustivité des menaces, ni efficacité de chaque contrôle pour chaque système. L'absence de contournement pendant un test n'est pas une preuve d'invulnérabilité. Les familles « misinformation » et « data poisoning » recoupent parfois des risques de qualité de l'information : leur traitement doit articuler sécurité, fiabilité, provenance et contrôle métier, sans tout assimiler à une cyberattaque.

La sécurité ne repose pas non plus exclusivement sur le modèle ou le fournisseur : elle dépend de l'architecture et de l'exploitation. Une application sécurisée aujourd'hui peut devenir vulnérable lorsque ses connecteurs, sa base documentaire ou ses droits changent.

08 / LIMITES — SUITE

Les références OWASP évoluent ; leurs éditions doivent être citées précisément. Les recommandations de ce rapport synthétisent des sources externes, elles ne décrivent pas une certification. Aucune attaque contre un produit réel, aucun taux de vulnérabilité et aucun benchmark propriétaire Spotwork.AI ne sont présentés. Les exemples de défense sont des modèles conceptuels à adapter à l'architecture et aux enjeux de chaque organisation.

L'efficacité des mécanismes d'isolation dépend des implémentations et des hypothèses de menace. Le filtrage de prompts, une consigne système renforcée, un « LLM-as-a-judge » ou une validation humaine peuvent contribuer à une défense en profondeur, mais ne remplacent pas des autorisations techniques et des contrôles d'accès robustes. Les journaux eux-mêmes doivent être sécurisés et ne pas devenir une nouvelle voie de divulgation.

09 Conclusion : sécuriser l'autonomie, pas seulement la réponse

En 2026, la question prioritaire n'est pas « notre LLM résiste-t-il aux mauvais prompts ? », mais **« que peut réellement lire, décider et exécuter notre système — et quelles preuves avons-nous que ces limites tiennent ? »**. Les guides OWASP éclairent les menaces ; le travail de l'entreprise consiste à les traduire en autorisations, contrôles, tests et responsabilités. C'est la condition d'une IA professionnelle gouvernable, même lorsque les modèles et les outils évoluent.

Implications pour les dirigeants

Le principal indicateur n'est pas le nombre de risques OWASP « cochés », mais la capacité à démontrer que les usages les plus sensibles restent dans des limites définies. Avant d'autoriser un agent, demander une vue de ses permissions, des données auxquelles il a accès, des tests négatifs, des mécanismes d'arrêt et des risques résiduels. Répéter cette décision lorsque changent les outils, les données ou le modèle.

10 Sources principales

[1] OWASP GenAI Security Project, OWASP Top 10 for LLM Applications 2026 (août 2026).
genai.owasp.org/resource/owasp-genai-llm-top-10-2026/

[2] OWASP GenAI Security Project, dépôt canonique de la liste 2026.
github.com/GenAI-Security-Project/GenAI-LLM-Top10/blob/main/2026/README.md

[3] OWASP GenAI Security Project, Top 10 for Agentic Applications for 2026 (décembre 2025).
genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

[4] OWASP GenAI Security Project, Agent Control Standard (septembre 2026).
genai.owasp.org/resource/agent-control-standard-acs/

[5] NIST, AI 100-2e2025 — Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (mars 2025).
nist.gov/publications/adversarial-machine-learning-taxonomy-and-terminology-attacks-and-mitigations-0

[6] NSA AISC, CISA, FBI, ASD ACSC, CCCS, NCSC-NZ et NCSC-UK, Deploying AI Systems Securely (avril 2024).
cyber.gov.au/publication/deploying-ai-systems-securely

[7] Cloud Security Alliance, note de recherche sur l'OWASP Top 10 LLM 2026 et l'Agent Control Standard.
labs.cloudsecurityalliance.org/research/csa-research-note-owasp-genai-top10-2026-agent-control-stand/

QUALIFICATION

Analyse documentaire sourcée et cadre méthodologique proposé par Spotwork.AI ; ce cadre n'est ni une extension officielle d'OWASP ni une mesure issue d'une expérimentation. Aucune validation de sécurité réelle, certification ou audit de produit ne peut en être déduit. Sources consultées au 2 novembre 2026.

Spotwork.AI

L'IA professionnelle de confiance

Mieux travailler aujourd'hui,
construire votre intelligence de demain.

SPOTLAB · RESEARCH & INSIGHTS · ANALYSE P09

Sécurité de l'IA générative en 2026 : ce que révèle OWASP et comment agir

Cinq frontières de confiance vérifiables

© 2026 SPOTWORK.AI