



# LLM AI Cybersecurity & Governance Checklist

From the OWASP Top 10  
for LLM Applications Team

**Version: 1.0**

*Published: February 19, 2024*

## Revision History

| Revision | Date       | Author(s)  | Description            |
|----------|------------|------------|------------------------|
| 0.1      | 2023-11-01 | Sandy Dunn | Draft initial          |
| 0.5      | 2023-12-06 | SD, Equipe | Draft public           |
| 0.9      | 2023-02-15 | SD, Equipe | Draft pré-version      |
| 1.0      | 2024-02-19 | SD, Equipe | version publique v 1.0 |

*Les informations fournies dans ce document ne constituent pas et ne sont pas destinées à constituer un avis juridique. Toutes les informations sont uniquement à des fins d'information générale.*

*Ce document contient des liens vers d'autres sites Web tiers. Ces liens sont uniquement destinés à des fins de commodité et l'OWASP ne recommande ni n'approuve le contenu des sites tiers.*

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Aperçu .....</b>   | <b>5</b>  |
| 1.1      | L'intelligence artificielle responsable et digne de confiance .....                           | 6         |
| 1.2      | À qui est-ce destiné ? .....  | 6         |
| 1.3      | Pourquoi une check-list ? .....   | 7         |
| 1.4      | Pas complet .....   | 7         |
| 1.5      | Défis des grands modèles de langage .....   | 7         |
| 1.6      | Catégories de menaces LLM .....   | 8         |
| 1.7      | Formation sur la sécurité et la confidentialité de l'intelligence artificielle .....          | 8         |
| 1.8      | Intégrer la sécurité et la gouvernance du LLM aux pratiques et contrôles existants et établis | 9         |
| 1.9      | Principes fondamentaux de sécurité .....  | 9         |
| 1.10     | Risques .....   | 9         |
| 1.11     | Taxonomie de vulnérabilité et d'atténuation .....   | 10        |
| <b>2</b> | <b>Déterminer la stratégie LLM .....</b>  | <b>11</b> |
| 2.1      | Stratégie de déploiement .....  | 13        |
| <b>3</b> | <b>Liste de contrôles .....</b>   | <b>14</b> |
| 3.1      | Risque contradictoire .....   | 14        |
| 3.2      | Modélisation des menaces .....  | 14        |
| 3.3      | Inventaire des actifs d'IA .....  | 15        |
| 3.4      | Formation sur la sécurité et la confidentialité de l'IA .....                                 | 16        |
| 3.5      | Établir des analyses de rentabilisation .....   | 16        |
| 3.6      | Gouvernance .....   | 17        |
| 3.7      | Légal .....   | 18        |
| 3.8      | Réglementaire .....   | 20        |
| 3.9      | Utilisation ou implémentation de solutions de grands modèles de langage .....                 | 21        |
| 3.10     | Tests, évaluation, vérification et validation (TEVV) .....                                    | 21        |
| 3.11     | Cartes modèles et cartes de risques .....   | 22        |
| 3.12     | RAG: optimisation des grands modèles de langage .....   | 23        |
| 3.13     | L'équipe Red Team IA .....  | 23        |

**4     Resources ..... 25**

**A     Team ..... 35**

# Aperçu

Chaque utilisateur d'Internet et chaque entreprise devraient se préparer à la prochaine vague d'applications puissantes d'intelligence artificielle générative ou IA générative (IAg ou GenAI). La GenAI est extrêmement prometteur en matière d'innovation, d'efficacité et de réussite commerciale dans une variété d'industries. Pourtant, comme toute technologie puissante à un stade précoce, elle comporte son propre ensemble de défis évidents et inattendus.

L'intelligence artificielle a considérablement progressé au cours des 50 dernières années, prenant discrètement en charge une variété de processus d'entreprise jusqu'à ce que l'apparition publique de ChatGPT stimule le développement et l'utilisation de grands modèles linguistiques parmi les particuliers et les entreprises. Initialement, ces technologies étaient limitées aux études universitaires ou à l'exécution d'activités certaines, mais vitales, au sein des entreprises, visibles uniquement par quelques privilégiés. Cependant, les progrès récents en matière de disponibilité des données, de puissance informatique, de capacités GenAI et de publication d'outils tels que Llama 2, ElevenLabs et Midjourney ont fait passer l'IA d'une niche à une acceptation généralisée. Ces améliorations ont non seulement rendu les technologies GenAI plus accessibles, mais elles ont également souligné la nécessité cruciale pour les entreprises de développer des stratégies solides pour intégrer et exploiter l'IA dans leurs opérations, ce qui représente un énorme pas en avant dans la façon dont nous utilisons la technologie.

- **Intelligence artificielle (IA)** est un terme général qui englobe tous les domaines de l'informatique permettant aux machines d'accomplir des tâches qui nécessiteraient normalement l'intelligence humaine. L'apprentissage automatique et l'IA générative sont deux sous-catégories de l'IA.
- **Apprentissage automatique ou Machine learning** est un sous-ensemble de l'IA qui se concentre sur la création d'algorithmes capables d'apprendre à partir des données. Les algorithmes d'apprentissage automatique sont formés sur un ensemble de données, puis peuvent utiliser ces données pour faire des prédictions ou prendre des décisions concernant de nouvelles données.
- L' **IA générative** est un type d'apprentissage automatique qui se concentre sur la création de nouvelles données.
- Un **grand modèle de langage (LLM pour "Large Language Models" en anglais)** est un type de modèle d'IA qui traite et génère du texte de type humain. Dans le contexte de l'intelligence artificielle, un « modèle » fait référence à un système formé pour faire des prédictions basées sur des données d'entrée. Les LLM sont spécifiquement formés sur de grands ensembles de données de langage naturel et sont appelés grands modèles de langage.

Les organisations pénètrent dans des territoires inexplorés en matière de sécurisation et de supervision des solutions GenAI. Les progrès rapides de GenAI permettent également aux adversaires d'améliorer leurs stratégies d'attaque, introduisant ainsi un double défi de défense et d'escalade des menaces.

Les entreprises utilisent l'intelligence artificielle dans de nombreux domaines, notamment les RH pour le recrutement, la détection du spam par courrier électronique, le SIEM pour l'analyse comportementale et les applications gérées de détection et de réponse. Cependant, ce document se concentre principalement sur les applications Large Language Model et leur fonction dans la création de contenu généré.

## L'intelligence artificielle responsable et digne de confiance

À mesure que les défis et les avantages de l'intelligence artificielle émergent - et que des réglementations et des lois sont adoptées - les principes et les piliers d'une utilisation responsable et digne de confiance de l'IA évoluent d'objets et de préoccupations idéalistes vers des normes établies. Le groupe de travail OWASP AI Exchange surveille ces changements et aborde les considérations plus larges et plus difficiles pour tous les aspects de l'intelligence artificielle.

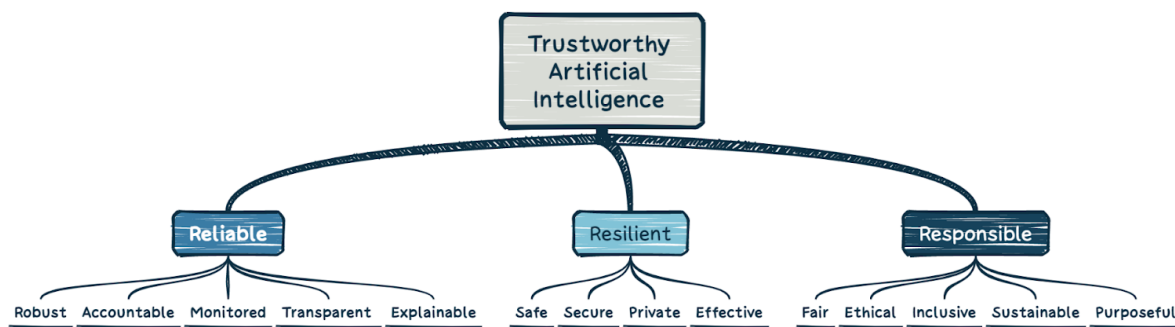


Figure 1.1: Image illustrant les piliers d'une intelligence artificielle fiable

## À qui est-ce destiné ?

La liste de contrôle OWASP Top 10 pour les applications LLM en matière de cybersécurité et de gouvernance s'adresse aux dirigeants des domaines de la direction, de la technologie, de la cybersécurité, de la confidentialité, de la conformité et du droit, aux équipes et défenseurs de DevSecOps, MLSecOps et de cybersécurité. Il est destiné aux personnes qui s'efforcent de garder une longueur d'avance dans le monde en évolution rapide de l'IA, dans le but non seulement d'exploiter l'IA pour le succès de l'entreprise, mais également de se protéger contre les risques de mises en œuvre hâtives ou non sécurisées de l'IA. Ces dirigeants et équipes doivent créer des stratégies pour saisir les opportunités, relever les défis et atténuer les risques.

Cette liste de contrôles est destinée à aider ces responsables technologiques et commerciaux à comprendre rapidement les risques et les avantages de l'utilisation du LLM, leur permettant de se concentrer sur l'élaboration d'une liste complète des domaines et des tâches critiques nécessaires pour défendre et protéger l'organisation lors de l'élaboration d'une stratégie de modèle linguistique étendu.

Le Top 10 OWASP pour l'équipe Applications LLM espère que cette liste aidera les organisations à améliorer leurs techniques défensives existantes et à développer des techniques pour faire face aux nouvelles menaces liées à l'utilisation de cette technologie passionnante.

## **Pourquoi une check-list ?**

Les listes de contrôles utilisées pour formuler des stratégies améliorent la précision, définissent les objectifs, préservent l'uniformité et favorisent un travail délibéré ciblé, réduisant ainsi les oublis et les détails manqués. Suivre une liste de contrôles augmente non seulement la confiance dans un parcours d'adoption sûr, mais encourage également les innovations des futures organisations en fournissant une stratégie simple et efficace d'amélioration continue.

## **Pas complet**

Bien que ce document vise à aider les organisations à développer une stratégie LLM initiale dans un environnement technique, juridique et réglementaire en évolution rapide, il n'est pas exhaustif et ne couvre pas tous les cas d'utilisation ou obligations. Lors de l'utilisation de ce document, les organisations doivent étendre les évaluations et les pratiques au-delà de la portée de la liste de contrôle fournie, selon les besoins de leur cas d'utilisation ou de leur juridiction.

## **Défis des grands modèles de langage**

Les grands modèles de langage sont confrontés à plusieurs problèmes sérieux et uniques. L'un des plus importants est que lorsque l'on travaille avec des LLM, les plans de contrôle et de données ne peuvent pas être strictement isolés ou séparables. Un autre défi important est que les LLM sont non déterministes de par leur conception, produisant un résultat différent lorsqu'ils y sont invités ou demandés.

Les LLM utilisent la recherche sémantique plutôt que la recherche par mots clés. La principale distinction entre les deux réside dans le fait que l'algorithme du modèle donne la priorité aux termes dans sa réponse. Il s'agit d'un changement important par rapport à la manière dont les consommateurs utilisaient auparavant la technologie, et cela a un impact sur la cohérence et la fiabilité des résultats. Les hallucinations, émergeant des lacunes et des défauts de formation dans les données sur lesquelles le modèle est formé, sont le résultat de cette méthode.

Il existe des méthodes pour améliorer la fiabilité et réduire la surface d'attaque pour le jailbreak, le model tricking et les hallucinations, mais il existe un compromis entre les restrictions et l'utilité, à la fois en termes de coût et de fonctionnalité.

L'utilisation du LLM et les applications LLM augmentent la surface d'attaque d'une organisation. Certains risques associés aux LLM sont uniques, mais beaucoup sont des problèmes familiers, tels que la nomenclature logicielle connue (SBOM pour "software bill of materials" en anglais), la chaîne d'approvisionnement, la protection contre la perte de données (DLP pour "data loss protection" en anglais) et l'accès autorisé. Il existe également des risques accrus qui ne sont pas directement liés à la GenAI, mais la GenAI augmente l'efficacité, la capacité et l'efficacité des attaquants qui attaquent et menacent les organisations.

Les adversaires exploitent de plus en plus les outils LLM et d'IA générative pour affiner et accélérer les méthodes traditionnelles d'attaque des organisations, des individus et des systèmes gouvernementaux.

Le LLM facilite leur capacité à améliorer les techniques leur permettant de créer sans effort de nouveaux logiciels malveillants, potentiellement intégrés à de nouvelles vulnérabilités zero-day ou conçus pour échapper à la détection. Ils peuvent également générer des programmes de phishing sophistiqués, uniques ou personnalisés. La création de deep fakes convaincants, qu'ils soient vidéo ou audio, favorise encore davantage leurs stratagèmes d'ingénierie sociale. De plus, ces outils leur permettent d'exécuter des intrusions et de développer des capacités de piratage innovantes. À l'avenir, une utilisation plus « adaptée » et plus complexe de la technologie de l'IA par les acteurs criminels exigera des réponses spécifiques et des solutions dédiées pour les capacités de défense et de résilience appropriées d'une organisation.

Les organisations sont également confrontées à la menace de NE PAS utiliser les capacités des LLM, telles qu'un désavantage concurrentiel, la perception du marché par les clients et les partenaires comme étant obsolètes, l'incapacité à faire évoluer les communications personnalisées, la stagnation de l'innovation, l'inefficacité opérationnelle, le risque plus élevé d'erreur humaine dans les processus, et allocation inefficace des ressources humaines.

Comprendre les différents types de menaces et les intégrer à la stratégie commerciale aidera à peser les avantages et les inconvénients de l'utilisation des grands modèles de langage par rapport à leur non-utilisation, en s'assurant qu'ils accélèrent plutôt qu'entravent la réalisation des objectifs commerciaux de l'entreprise.

## Catégories de menaces LLM



Figure 1.2: Image illustrant les types de menaces liées à l'IA

## Formation sur la sécurité et la confidentialité de l'intelligence artificielle

Les employés de toutes les organisations bénéficient d'une formation pour comprendre l'intelligence artificielle, l'intelligence artificielle générative et les conséquences potentielles futures de la création, de l'achat ou de l'utilisation de LLM. La formation aux utilisations autorisées et à la sensibilisation à la sécurité devrait cibler tous les employés et être plus spécialisée pour certains postes tels que les ressources humaines, les services juridiques, les développeurs, les équipes de données et les équipes de sécurité.

Les politiques d'utilisation équitable et les interactions saines sont des aspects clés qui, s'ils sont intégrés dès le début, constitueront la pierre angulaire du succès des futures campagnes



de sensibilisation à la cybersécurité de l'IA. Cela fournira nécessairement aux utilisateurs une connaissance des règles de base de l'interaction ainsi que la capacité de distinguer les bons comportements des mauvais comportements ou des comportements contraires à l'éthique.

## **Intégrer la sécurité et la gouvernance du LLM aux pratiques et contrôles existants et établis**

Même si l'IA et l'IA générée ajoutent une nouvelle dimension à la cybersécurité, à la résilience, à la confidentialité et au respect des exigences légales et réglementaires, les meilleures pratiques qui existent depuis longtemps restent le meilleur moyen d'identifier les problèmes, de trouver les vulnérabilités, de les corriger et de les résoudre. atténuer les problèmes de sécurité potentiels.

- Confirmer que la gestion des systèmes d'intelligence artificielle est intégrée aux pratiques organisationnelles existantes.
- Confirmer que les systèmes AIML ("langage de balisage d'intelligence artificielle" ou "Artificial Intelligence Markup Language", en anglais) suivent les pratiques existantes en matière de confidentialité, de gouvernance et de sécurité, avec des pratiques de confidentialité, de gouvernance et de sécurité spécifiques à l'IA mises en œuvre si nécessaire.

## **Principes fondamentaux de sécurité**

Les capacités LLM introduisent un type différent d'attaque et de surface d'attaque. Les LLM sont vulnérables aux bogues de logique métier complexes, tels que l'injection rapide, la conception de plugins non sécurisées et l'exécution de code à distance. Les meilleures pratiques existantes constituent le meilleur moyen de résoudre ces problèmes. Une équipe interne de sécurité des produits qui comprend l'examen des logiciels sécurisés, l'architecture, la gouvernance des données et les évaluations tierces. L'équipe de cybersécurité doit également vérifier la solidité des contrôles actuels pour détecter les problèmes qui pourraient être aggravés par LLM, tels que le clonage vocal, l'usurpation d'identité, ou en contournant les captchas.

Compte tenu des progrès récents en matière d'apprentissage automatique, de NLP (Traitement automatique des langues ou "Natural Language Processing" en anglais), de NLU (Compréhension du langage naturel ou "Natural Language Understanding" en anglais), de Deep Learning et, plus récemment, de LLM et d'IA générative, il est recommandé d'inclure des professionnels compétents dans ces domaines aux côtés de la cybersécurité et équipes devops. Leur expertise contribuera non seulement à l'adoption de ces technologies, mais également à l'élaboration d'analyses et de réponses innovantes aux défis émergents.

## **Risques**

La référence au risque utilise la définition ISO 31000 : Risque = « effet de l'incertitude sur les objectifs ». Les risques LLM inclus dans la liste de contrôle comprennent une liste ciblée de risques LLM qui traitent des risques contradictoires, de sécurité, juridiques, réglementaires, de réputation, financiers et concurrentiels.

## **Taxonomie de vulnérabilité et d'atténuation**

Les systèmes actuels de classification des vulnérabilités et de partage d'informations sur les menaces, comme OVAL, STIX, CVE et CWE, développent encore la capacité de surveiller et d'alerter les défenseurs des vulnérabilités et des menaces spécifiques aux grands modèles de langage et aux modèles prédictifs. On s'attend à ce que les organisations s'appuient sur ces normes établies et reconnues, telles que CVE pour la classification des vulnérabilités et STIX pour l'échange de renseignements sur les cybermenaces (CTI), lorsque des vulnérabilités ou des menaces pesant sur les systèmes d'IA/ML et leurs chaînes d'approvisionnement sont identifiées.

# Déterminer la stratégie LLM

L'expansion rapide des applications LLM a accru l'attention et l'examen de tous les systèmes d'IA/ML utilisés dans les opérations commerciales, englobant à la fois l'IA générative et les systèmes d'IA/ML prédictifs établis de longue date. Cette attention accrue expose des risques potentiels, tels que des attaquants ciblant des systèmes qui étaient auparavant négligés et des problèmes de gouvernance ou juridiques qui auraient pu être ignorés en termes de problèmes juridiques, de confidentialité, de responsabilité ou de garantie. Pour toute organisation tirant parti des systèmes d'IA/ML dans ses opérations, il est essentiel d'évaluer et d'établir des politiques complètes, une gouvernance, des protocoles de sécurité, des mesures de confidentialité et des normes de responsabilité pour garantir que ces technologies s'alignent sur les processus commerciaux de manière sécurisée et éthique.

Les attaquants, ou adversaires, constituent la menace la plus immédiate et la plus nuisible pour les entreprises, les personnes et les agences gouvernementales. Leurs objectifs, qui vont du gain financier à l'espionnage, les poussent à voler des informations critiques, à perturber les opérations et à nuire à la confiance. De plus, leur capacité à exploiter de nouvelles technologies telles que l'IA et l'apprentissage automatique augmente la vitesse et la sophistication des attaques, ce qui rend difficile pour les défenses de garder une longueur d'avance sur les attaques.

La menace LLM non adverse la plus pressante pour de nombreuses organisations provient de « Shadow AI » : des employés utilisant des outils d'IA en ligne non approuvés, des plug-ins de navigateur dangereux et des applications tierces qui introduisent des fonctionnalités LLM via des mises à jour ou des mises à niveau, contournant les processus d'approbation de logiciels standard.

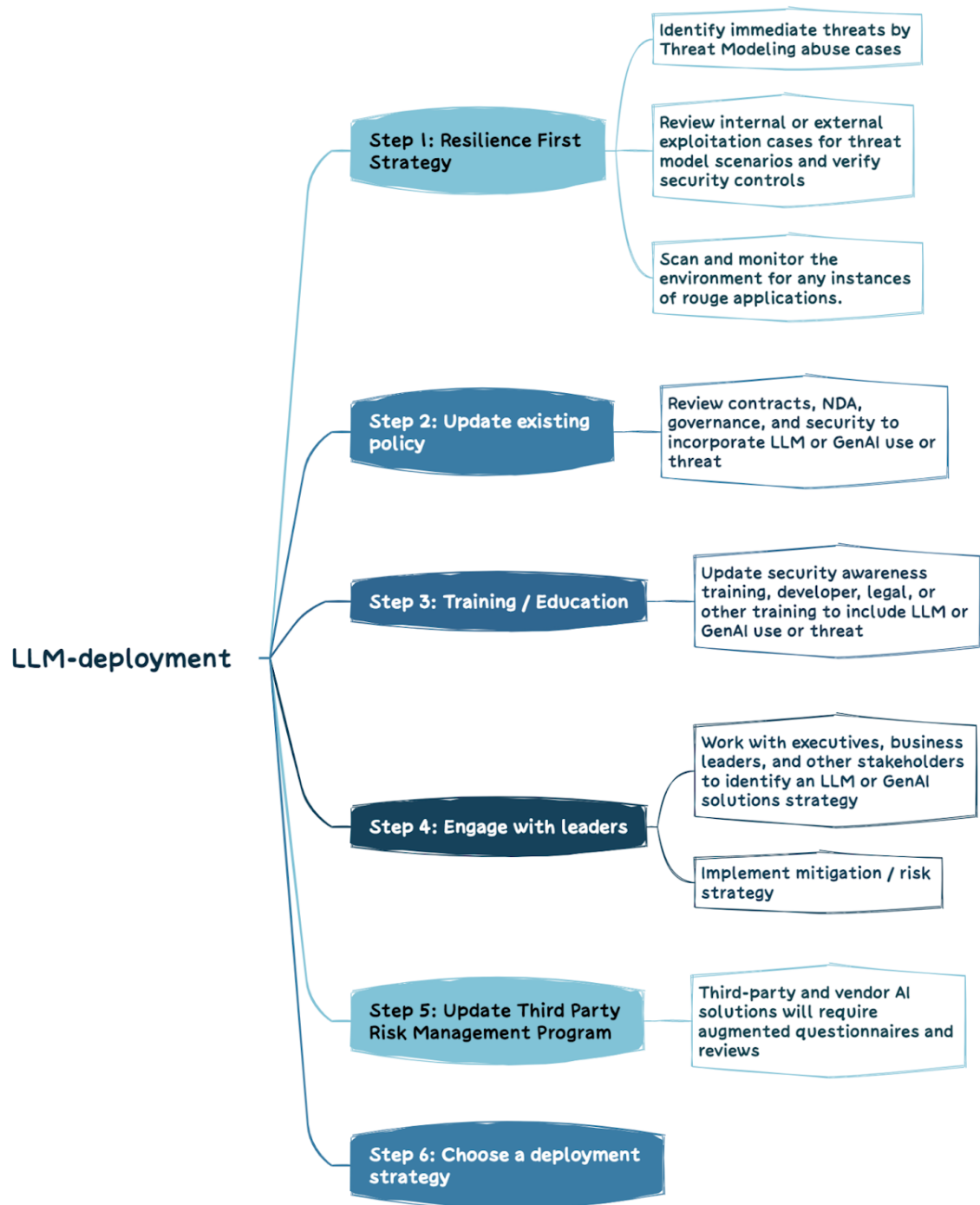


Figure 2.1: Image des options pour la stratégie de déploiement

## Stratégie de déploiement

Les domaines d'application vont de l'exploitation d'applications grand public à la formation de modèles propriétaires sur des données privées. Des facteurs tels que la sensibilité des cas d'utilisation, les capacités nécessaires et les ressources disponibles aident à déterminer le bon équilibre entre commodité et contrôle. Cependant, la compréhension de ces cinq types de modèles fournit un cadre pour évaluer les options.

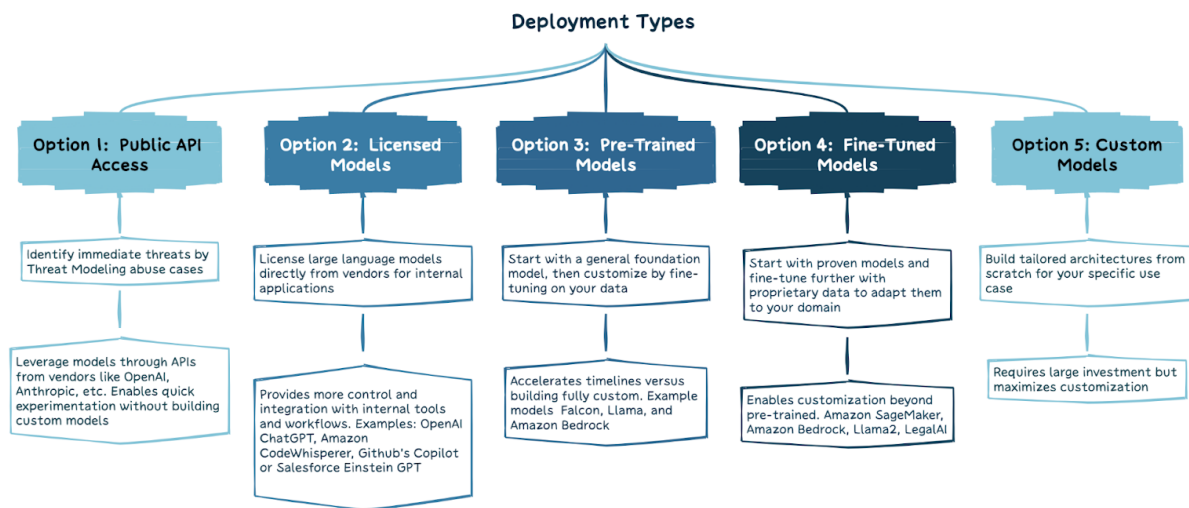


Figure 2.2: Image des options pour les types de déploiement

# Liste de contrôles

## Risque contradictoire

Le risque contradictoire inclut les concurrents et les attaquants.

- ❑ Examiner la manière dont les concurrents investissent dans l'intelligence artificielle. Bien que l'adoption de l'IA comporte des risques, elle présente également des avantages commerciaux qui peuvent avoir un impact sur les positions futures sur le marché.
- ❑ Étudiez l'impact des contrôles actuels, tels que la réinitialisation des mots de passe, qui utilisent la reconnaissance vocale et qui pourraient ne plus fournir la sécurité défensive appropriée contre les nouvelles attaques améliorées de la GenAI.
- ❑ Mettez à jour le plan de réponse aux incidents et les playbooks pour les attaques améliorées de la GenAI et les incidents spécifiques à l'AIML.

## Modélisation des menaces

La modélisation des menaces est fortement recommandée pour identifier les menaces et examiner les processus et les défenses de sécurité. La modélisation des menaces est un ensemble de processus systématiques et reproductibles qui permettent de prendre des décisions raisonnables en matière de sécurité pour les applications, les logiciels et les systèmes. La modélisation des menaces pour les attaques accélérées de la GenAI et avant le déploiement du LLM est le moyen le plus rentable d'identifier et d'atténuer les risques, de protéger les données, de protéger la confidentialité et d'assurer une intégration sécurisée et conforme au sein de l'entreprise.

- ❑ Comment les attaquants vont-ils accélérer les attaques contre l'organisation, les employés, les dirigeants ou les utilisateurs ? Les organisations doivent anticiper les attaques « hyper-personnalisées » à grande échelle grâce à l'IA générative. Les attaques de spear phishing assistées par LLM sont désormais exponentiellement plus efficaces, ciblées et militarisées pour une attaque.
- ❑ Comment la GenAI pourrait-il être utilisé pour des attaques contre les clients de l'entreprise via l'usurpation d'identité ou le contenu généré par la GenAI ?
- ❑ L'entreprise peut-elle détecter et neutraliser les entrées ou requêtes nuisibles ou malveillantes adressées aux solutions LLM ?
- ❑ L'entreprise peut-elle protéger les connexions avec les systèmes et bases de données existants avec des intégrations sécurisées à toutes les limites de confiance LLM ?
- ❑ L'entreprise dispose-t-elle d'une atténuation des menaces internes pour empêcher toute utilisation abusive par les utilisateurs autorisés ?
- ❑ L'entreprise peut-elle empêcher l'accès non autorisé à des modèles ou des données propriétaires pour protéger la propriété intellectuelle ?
- ❑ L'entreprise peut-elle empêcher la génération de contenu nuisible ou inapproprié grâce au filtrage automatisé du contenu ?

## **Inventaire des actifs d'IA**

Un inventaire des actifs d'IA doit s'appliquer à la fois aux solutions développées en interne et aux solutions externes ou tierces.

- ❑ Cataloguer les services, outils et propriétaires d'IA existants. Désigner une balise dans la gestion des actifs pour un inventaire spécifique.
- ❑ Inclure les composants IA dans la nomenclature logicielle (SBOM pour "Software Bill of Materials" en anglais), une liste complète de tous les composants logiciels, dépendances et métadonnées associés aux applications.
- ❑ Cataloguer les sources de données IA et la sensibilité des données (protégées, confidentielles, publiques)
- ❑ Déterminer si des tests d'intrusion ou une équipe rouge des solutions d'IA déployées sont nécessaires pour déterminer le risque actuel de surface d'attaque.
- ❑ Créer un processus d'intégration de solution d'IA.
- ❑ S'assurer que le personnel administratif informatique qualifié est disponible en interne ou en externe, conformément aux exigences SBoM.

## **Formation sur la sécurité et la confidentialité de l'IA**

- ❑ S'engager activement avec les employés pour comprendre et répondre aux préoccupations concernant les initiatives LLM prévues.
- ❑ Établir une culture de communication ouverte et transparente sur l'utilisation par l'organisation de l'IA prédictive ou générative au sein des processus, des systèmes, de la gestion et du support des employés, des engagements des clients et sur la manière dont son utilisation est régie, gérée et les risques gérés.
- ❑ Former tous les utilisateurs à l'éthique, à la responsabilité et aux questions juridiques telles que la garantie, la licence et le droit d'auteur.
- ❑ Mettre à jour la formation de sensibilisation à la sécurité pour inclure les menaces liées à la GenAI. Clonage de voix et clonage d'images, ainsi qu'en prévision de l'augmentation des attaques de spear phishing
- ❑ Toute solution GenAI adoptée devrait inclure une formation aux DevOps et à la cybersécurité pour le pipeline de déploiement afin de garantir la sûreté et la sécurité de l'IA.

## **Établir des analyses de rentabilisation**

Des analyses de rentabilisation solides sont essentielles pour déterminer la valeur commerciale de toute solution d'IA proposée, équilibrer les risques et les avantages, et évaluer et tester le retour sur investissement. Il existe un très grand nombre de cas d'utilisation potentiels ; quelques exemples sont fournis.

- ❑ Améliorer l'expérience client
- ❑ Meilleure efficacité opérationnelle
- ❑ Une meilleure gestion des connaissances
- ❑ Innovation améliorée
- ❑ Études de marché et analyse des concurrents
- ❑ Création, traduction, synthèse et analyse de documents



## Gouvernance

La gouvernance d'entreprise en LLM est nécessaire pour fournir aux organisations transparence et responsabilité. L'identification des propriétaires de plates-formes ou de processus d'IA potentiellement familiers avec la technologie ou les cas d'utilisation sélectionnés pour l'entreprise est non seulement conseillé, mais également nécessaire pour garantir une vitesse de réaction adéquate qui évite les dommages collatéraux aux processus numériques d'entreprise bien établis.

- ❑ Établir la matrice RACI IA de l'organisation (qui est responsable, qui doit rendre des comptes, qui doit être consulté et qui doit être informé)
- ❑ Documentez et attribuez les risques liés à l'IA, les évaluations des risques et les responsabilités de gouvernance au sein de l'organisation.
- ❑ Établir des politiques de gestion des données, y compris l'application technique, concernant la classification des données et les limitations d'utilisation. Les modèles ne doivent exploiter que les données classées pour le niveau d'accès minimum de tout utilisateur du système. Par exemple, mettez à jour la politique de protection des données pour insister sur le fait de ne pas saisir de données protégées ou confidentielles dans des outils non gérés par l'entreprise.
- ❑ Créer une politique d'IA soutenue par une politique établie (par exemple, norme de bonne conduite, protection des données, utilisation des logiciels)
- ❑ Publier une matrice d'utilisation acceptable pour divers outils d'IA générative que les employés pourront utiliser.
- ❑ Documenter les sources et la gestion de toutes les données que l'organisation utilise à partir des modèles LLM génératifs.

## **Légal**

De nombreuses implications juridiques de l'IA restent floues et potentiellement très coûteuses. Un partenariat informatique, sécuritaire et juridique est essentiel pour identifier les lacunes et prendre des décisions obscures.

- ❑ Confirmer que les garanties des produits sont claires dans le flux de développement de produits pour désigner qui est responsable des garanties des produits avec l'IA.
- ❑ Examinez et mettez à jour les termes et conditions existants pour toute considération relative à la GenAI.
- ❑ Consultez les accords CLUF sur l'IA. Les accords de licence d'utilisateur final pour les plates-formes GenAI sont très différents dans la manière dont ils traitent les invites des utilisateurs, les droits et la propriété des résultats, la confidentialité des données, la conformité, la responsabilité, la confidentialité et les limites d'utilisation des résultats.
- ❑ CLUF de l'organisation pour les clients, modifier les accords d'utilisateur final pour empêcher l'organisation d'encourir des responsabilités liées au plagiat, à la propagation de préjugés ou à la violation de la propriété intellectuelle via le contenu généré par l'IA.
- ❑ Passer en revue les outils assistés par l'IA existants utilisés pour le développement de code. La capacité d'un chatbot à écrire du code peut menacer les droits de propriété d'une entreprise sur son produit si un chatbot est utilisé pour générer du code pour le produit. Par exemple, cela pourrait remettre en question le statut et la protection du contenu généré et qui détient le droit d'utiliser le contenu généré.
- ❑ Examiner tous les risques liés à la propriété intellectuelle. La propriété intellectuelle générée par un chatbot pourrait être menacée si des données obtenues de manière inappropriée étaient utilisées au cours du processus de génération, qui est soumis à la protection des droits d'auteur, des marques ou des brevets. Si les produits d'IA utilisent du matériel contrefait, cela crée un risque pour les résultats de l'IA, ce qui peut entraîner une violation de la propriété intellectuelle.
- ❑ Examiner tous les contrats comportant des dispositions d'indemnisation. Les clauses d'indemnisation tentent de faire porter la responsabilité d'un événement entraînant une responsabilité sur celui qui en était le plus fautif ou qui avait le plus de chances de l'arrêter. Établir des garde-fous pour déterminer si le fournisseur de l'IA ou son utilisateur est à l'origine de l'événement, engageant sa responsabilité.
- ❑ Examiner la responsabilité pour les blessures potentielles et les dommages matériels causés par les systèmes d'IA.
- ❑ Examiner la couverture d'assurance. Les polices d'assurance responsabilité civile générale (D&O) et commerciale générale sont probablement insuffisantes pour protéger pleinement l'utilisation de l'IA.
- ❑ Identifier tout problème de droits d'auteur. La paternité humaine est requise pour le droit d'auteur. Une organisation peut également être tenue responsable de plagiat, de propagation de préjugés ou de violation de la propriété intellectuelle si les outils LLM sont utilisés à mauvais escient.
- ❑ Veiller à ce que des accords soient en place pour les sous-traitants et à l'utilisation appropriée de l'IA pour tout développement ou service fourni.
- ❑ Restreindre ou interdire l'utilisation d'outils d'IA générative pour les employés ou les sous-traitants lorsque les droits exécutoires peuvent poser problème ou lorsqu'il existe des problèmes de violation de la propriété intellectuelle.
- ❑ Les solutions d'évaluation et d'IA utilisées pour la gestion ou l'embauche des employés pourraient donner lieu à des demandes de traitement ou d'impact disparates.
- ❑ S'assurer que les solutions d'IA ne collectent ni ne partagent d'informations sensibles sans le consentement ou l'autorisation appropriée.

## Réglementaire

La loi européenne sur l'IA devrait être la première loi complète sur l'IA, mais elle s'appliquera au plus tôt en 2025. Le règlement général sur la protection des données (RGPD) de l'UE ne traite pas spécifiquement de l'IA, mais comprend des règles relatives à la collecte de données, à la sécurité des données, à l'équité et à la transparence, à l'exactitude et à la fiabilité, ainsi qu'à la responsabilité, qui peuvent avoir un impact sur l'utilisation de la GenAI. Aux États-Unis, la réglementation de l'IA est incluse dans des lois plus larges sur la protection de la vie privée des consommateurs. Dix États américains ont adopté ou disposent de lois qui entreront en vigueur d'ici la fin de 2023.

Des organisations fédérales telles que la Commission américaine pour l'égalité des chances en matière d'emploi (EEOC), le Bureau de protection financière des consommateurs (CFPB), la Commission fédérale du commerce (FTC) et la Division des droits civils (DOJ) du ministère américain de la Justice surveillent de près l'équité en matière d'embauche.

- ☐ Déterminer les exigences de conformité en matière d'IA spécifiques au pays, à l'État ou à tout autre gouvernement.
- ☐ Déterminer les exigences de conformité pour restreindre la surveillance électronique des employés et les systèmes de décision automatisés liés à l'emploi (Vermont, Californie, Maryland, New York, New Jersey)
- ☐ Déterminer les exigences de conformité en matière de consentement à la reconnaissance faciale et à l'analyse vidéo IA requise (Illinois, Maryland, Washington, Vermont)
- ☐ Passer en revue tous les outils d'IA utilisés ou envisagés pour l'embauche ou la gestion des employés.
- ☐ Confirmer la conformité du fournisseur aux lois et aux meilleures pratiques applicables en matière d'IA.
- ☐ Demandez et documentez tous les produits utilisant l'IA pendant le processus d'embauche. Demandez comment le modèle a été formé et comment il est surveillé, et suivez toutes les corrections apportées pour éviter la discrimination et les préjugés.
- ☐ Demander et documenter quelles options d'hébergement sont incluses.
- ☐ Demander et documenter si le fournisseur collecte des données confidentielles.
- ☐ Demander comment le fournisseur ou l'outil stocke et supprime les données et réglemente l'utilisation des outils de reconnaissance faciale et d'analyse vidéo pendant la période préalable à l'emploi.
- ☐ Examiner les autres exigences réglementaires spécifiques à l'organisation avec l'IA qui peuvent soulever des problèmes de conformité. La Loi sur la sécurité du revenu de retraite des employés de 1974, par exemple, impose des obligations fiduciaires pour les régimes de retraite qu'un chatbot pourrait ne pas être en mesure de respecter.

## Utilisation ou implémentation de solutions de grands modèles de langage

- ❑ Composants LLM du modèle de menace et limites de confiance de l'architecture.
- ❑ Sécurité des données : vérifier comment les données sont classées et protégées en fonction de leur sensibilité, y compris les données professionnelles personnelles et exclusives. (Comment les autorisations des utilisateurs sont-elles gérées et quelles protections sont en place?)
- ❑ Contrôle d'accès : mise en œuvre de contrôles d'accès au moindre privilège et mise en œuvre de mesures de défense en profondeur
- ❑ La sécurité des pipelines de formation nécessite un contrôle rigoureux de la gouvernance des données de formation, des pipelines, des modèles et des algorithmes.
- ❑ Sécurité des entrées et des sorties : évaluer les méthodes de validation des entrées, ainsi que la manière dont les sorties sont filtrées, nettoyées et approuvées.
- ❑ Surveillance et réponse : cartographier les flux de travail, la surveillance et les réponses pour comprendre l'automatisation, la journalisation et l'audit. Confirmer que les enregistrements d'audit sont sécurisés.
- ❑ Inclure les tests d'application, l'examen du code source, les évaluations de vulnérabilité et l'équipe Red Team dans le processus de publication de production.
- ❑ Rechercher les vulnérabilités existantes dans le modèle LLM ou la chaîne d'approvisionnement.
- ❑ Examiner les effets des menaces et des attaques sur les solutions LLM, telles que l'injection rapide, la divulgation d'informations sensibles et la manipulation de processus.
- ❑ Étudier l'impact des attaques et des menaces sur les modèles LLM, notamment l'empoisonnement des modèles, la gestion inappropriée des données, les attaques de la chaîne d'approvisionnement et le vol de modèles.
- ❑ Sécurité de la chaîne d'approvisionnement, demandez des audits tiers, des tests d'intrusion et des révisions de code pour les fournisseurs tiers (à la fois initialement et de manière continue).
- ❑ Sécurité de l'infrastructure, se demander à quelle fréquence un fournisseur effectue des tests de résilience? Quels sont leurs SLA en termes de disponibilité, d'évolutivité et de performances?
- ❑ Mettre à jour les playbooks de réponse aux incidents et incluez un incident LLM dans les exercices théoriques.
- ❑ Identifier ou développer des mesures pour comparer l'IA de cybersécurité générative à d'autres approches afin de mesurer les améliorations de productivité attendues.

## Tests, évaluation, vérification et validation (TEVV)

Le cadre NIST AI recommande un processus TEVV continu tout au long du cycle de vie de l'IA qui inclut les opérateurs du système d'IA, les experts du domaine, les concepteurs d'IA, les utilisateurs, les développeurs de produits, les évaluateurs et les auditeurs. TEVV comprend une gamme de tâches telles que la validation du système, l'intégration, les tests, le recalibrage et la surveillance continue pour des mises à jour périodiques afin de gérer les risques et les changements du système d'IA.

- ❑ Établir des tests, des évaluations, des vérifications et des validations continus tout au long du cycle de vie du modèle d'IA.
- ❑ Fournir des mesures exécutives régulières et des mises à jour sur la fonctionnalité, la sécurité, la fiabilité et la robustesse du modèle d'IA.

## Cartes modèles et cartes de risques

Les cartes modèles et les cartes de risque sont des éléments fondamentaux pour accroître la transparence, la responsabilité et le déploiement éthique des grands modèles linguistiques (LLM). Les cartes modèles aident les utilisateurs à comprendre et à faire confiance aux systèmes d'IA en fournissant une documentation standardisée sur leur conception, leurs capacités et leurs contraintes, les amenant ainsi à créer des applications informées et sûres. Les cartes de risque complètent cela en abordant ouvertement les conséquences négatives potentielles, telles que les préjugés, les problèmes de confidentialité et les vulnérabilités de sécurité, ce qui encourage une approche proactive de la prévention des dommages. Ces documents sont essentiels pour les développeurs, les utilisateurs, les régulateurs et les éthiciens, car ils établissent une atmosphère collaborative dans laquelle les implications sociales de l'IA sont soigneusement abordées et traitées. Ces cartes, développées et gérées par les organisations qui ont créé les modèles, jouent un rôle important en garantissant que les technologies d'IA respectent les normes éthiques et les exigences légales, permettant ainsi une recherche et un déploiement responsables dans l'écosystème de l'IA.

Les cartes de modèle incluent des attributs clés associés au modèle ML :

- **Détails du modèle:** Informations de base sur le modèle, c'est-à-dire son nom, sa version et son type (réseau neuronal, arbre de décision, etc.), ainsi que le cas d'utilisation prévu.
- **Architecture du modèle:** Comprend une description de la structure du modèle, telle que le nombre et le type de couches, les fonctions d'activation et d'autres choix architecturaux clés.
- **Données et méthodologie de formation:** Informations sur les données utilisées pour entraîner le modèle, telles que la taille de l'ensemble de données, les sources de données et les techniques de prétraitement ou d'augmentation des données utilisées. Il comprend également des détails sur la méthodologie de formation, tels que l'optimiseur utilisé, la fonction de perte et tous les hyperparamètres réglés.
- **Indicateurs de performance:** Informations sur les performances du modèle sur diverses mesures, telles que l'exactitude, la précision, le rappel et le score F1. Il peut également inclure des informations sur les performances du modèle sur différents sous-ensembles de données.
- **Biais et limites potentiels:** Répertoire des biais ou limitations potentiels du modèle, tels que des données d'entraînement déséquilibrées, un surajustement ou des biais dans les prédictions du modèle. Il peut également inclure des informations sur les limites du modèle, telles que sa capacité à se généraliser à de nouvelles données ou son adéquation à certains cas d'utilisation.
- **Considérations sur l'IA responsable:** Toute considération éthique ou responsable en matière d'IA liée au modèle, telle que les problèmes de confidentialité, d'équité et de transparence, ou les impacts sociétaux potentiels de l'utilisation du modèle. Il peut également inclure des recommandations pour des tests, une validation ou une surveillance plus approfondis du modèle.

Les fonctionnalités précises contenues dans une carte modèle peuvent différer en fonction du contexte du modèle et de l'utilisation prévue, mais l'objectif est de donner ouverture et responsabilité dans la création et le déploiement de modèles d'apprentissage automatique.

- ☐ Examiner une fiche de modèle de modèles
- ☐ Examiner la carte de risque si disponible
- ☐ Établir un processus pour suivre et conserver les fiches de modèle pour tout modèle déployé, y compris les modèles utilisés par un tiers.

## **RAG: optimisation des grands modèles de langage**

Le réglage fin, la méthode traditionnelle d'optimisation d'un modèle pré-entraîné, impliquait de recycler un modèle existant sur des données nouvelles et spécifiques à un domaine, en le modifiant pour les performances d'une tâche ou d'une application. Le réglage fin est coûteux mais essentiel pour améliorer les performances.

La génération augmentée par récupération (RAG pour "Retrieval-Augmented Generation" en anglais) a évolué pour devenir un moyen plus efficace d'optimiser et d'augmenter les capacités des grands modèles de langage en récupérant des données pertinentes à partir de sources de connaissances disponibles à jour. RAG peut être personnalisé pour des domaines spécifiques, optimisant la récupération d'informations spécifiques au domaine et adaptant le processus de génération aux nuances des domaines spécialisés. RAG est considéré comme une méthode plus efficace et transparente pour l'optimisation LLM, en particulier pour les problèmes où les données étiquetées sont limitées ou coûteuses à collecter. L'un des principaux avantages de RAG est sa prise en charge de l'apprentissage continu puisque les nouvelles informations peuvent être continuellement mises à jour au stade de la récupération.

La mise en œuvre du RAG implique plusieurs étapes clés, depuis l'intégration du déploiement du modèle, l'indexation de la bibliothèque de connaissances jusqu'à la récupération des documents les plus pertinents pour le traitement des requêtes. Une récupération efficace du contexte pertinent est effectuée sur la base de bases de données vectorielles qui sont utilisées pour le stockage et l'interrogation des intégrations de documents.

### **Référence RAG**

- ☐ Retrieval Augmented Generation (RAG) & LLM: Examples
- ☐ 12 RAG Pain Points and Proposed Solutions

## **L'équipe Red Team IA**

L'équipe Red Team IA est une simulation de test d'attaque contradictoire du système IA pour valider qu'il n'existe aucune vulnérabilité existante pouvant être exploitée par un attaquant. Il s'agit d'une pratique recommandée par de nombreux organismes de réglementation et de gouvernance de l'IA, y compris l'administration Biden. L'équipe rouge ne constitue pas à elle seule une solution complète pour valider tous les préjudices réels associés aux systèmes d'IA et doit être incluse avec d'autres formes de tests, d'évaluation, de vérification et de validation telles que les évaluations d'impact algorithmiques et les audits externes.

- Intégrer les tests Red Team comme pratique standard pour les modèles et applications d'IA.



# Resources

## OWASP Top 10 for Large Language Model Applications

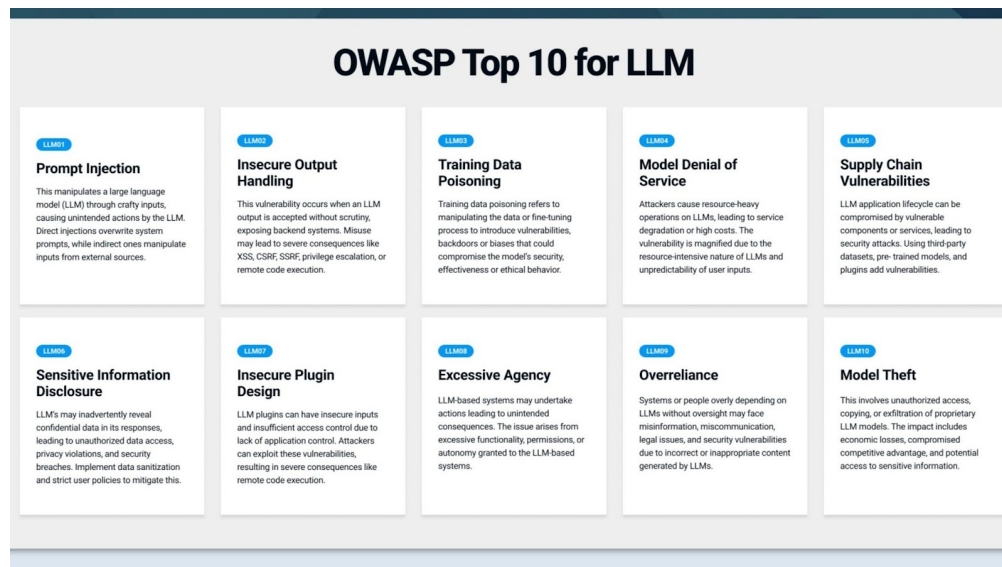


Figure 4.1: Image of OWASP Top 10 for Large Language Model Applications

## OWASP Top 10 for Large Language Model Applications Visualized

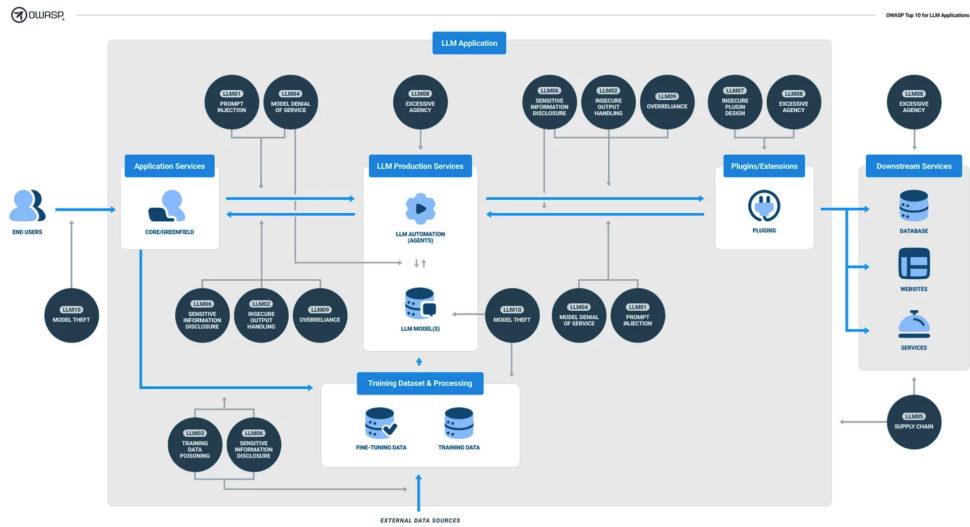


Figure 4.2: Image of OWASP Top 10 for Large Language Model Applications Visualized

**OWASP Resources** Using LLM solutions expands an organization's attack surface and presents new challenges, requiring special tactics and defenses. It also poses problems that are similar to known issues, and there are already established cybersecurity procedures and mitigations. Integrating LLM cybersecurity with an organization's established cybersecurity controls, processes, and procedures allows an organization to reduce its vulnerability to threats. How they integrate with each other is available at the OWASP Integration Standards.

| OWASP Resource                      | Description   | Why It Is Recommended & Where To Use It   |
|-------------------------------------|---|---|
| OWASP SAMM                          | Software Assurance Maturity Model   | Provides an effective and measurable way to analyze and improve an organization's secure development lifecycle. SAMM supports the complete software lifecycle. It is iterative and risk-driven, enabling organizations to identify and prioritize gaps in secure software development so resources for improving the process can be dedicated where efforts have the greatest improvement impact. |
| OWASP AI Security and Privacy Guide | OWASP Project with a goal of connecting worldwide for an exchange on AI security, fostering standards alignment, and driving collaboration. | The OWASP AI Security and Privacy Guide is a comprehensive list of the most important AI security and privacy considerations. It is meant to be a comprehensive resource for developers, security researchers, and security consultants to verify the security and privacy of AI systems.   |
| OWASP AI Exchange                   | OWASP AI Exchange is the intake method for the OWASP AI Security and Privacy Guide.   | The AI Exchange is the primary intake method used by OWASP to drive the direction of the OWASP AI Security and Privacy Guide.   |

| OWASP Resource                         | Description   | Why It Is Recommended & Where To Use It  |
|--|---|--|
| OWASP Machine Learning Security Top 10 | OWASP Machine Learning Security Top 10 security issues of machine learning systems.   | The OWASP Machine Learning Security Top 10 is a community-driven effort to collect and present the most important security issues of machine learning systems in a format that is easy to understand by both a security expert and a data scientist. This project includes the ML Top 10 and is a live working document that provides clear and actionable insights on designing, creating, testing, and procuring secure and privacy-preserving AI systems. It is the best OWASP resource for AI global regulatory and privacy information. |
| OpenCRE                                | OpenCRE (Common Requirement Enumeration) is the interactive content-linking platform for uniting security standards and guidelines into one overview. | Use this site to search for standards. You can search by standard name or by control type.   |
| OWASP Threat Modeling                  | A structured, formal process for threat modeling of an application  | Learn everything about Threat Modeling which is a structured representation of all the information that affects the security of an application.  |
| OWASP CycloneDX                        | OWASP CycloneDX is a full-stack Bill of Materials (BOM) standard that provides advanced supply chain capabilities for cyber risk reduction.           | Modern software is assembled using third-party and open source components. They are glued together in complex and unique ways and integrated with original code to achieve the desired functionality. An SBOM provides an accurate inventory of all components which enables organizations to identify risk, allows for greater transparency, and enables rapid impact analysis. EO 14028 provided minimum requirements for SBOM for federal systems.  |

| OWASP Resource  | Description   | Why It Is Recommended & Where To Use It  |
|---|---|--|
| OWASP Software Component Verification Standard (SCVS) | A community-driven effort to establish a framework for identifying activities, controls, and best practices can help in identifying and reducing risk in a software supply chain.                                   | Use SCVS to develop a common set of activities, controls, and best-practices that can reduce risk in a software supply chain and identify a baseline and path to mature software supply chain vigilance.   |
| OWASP API Security Project                            | API Security focuses on strategies and solutions to understand and mitigate the unique vulnerabilities and security risks of Application Programming Interfaces (APIs)  | APIs are a foundational element of connecting applications, and mitigating misconfigurations or vulnerabilities is mandatory to protect users and organizations. Use for security testing and red teaming the build and production environments.                           |
| OWASP Application Security Verification Standard ASVS | Application Security Verification Standard (ASVS) Project provides a basis for testing web application technical security controls and also provides developers with a list of requirements for secure development. | Cookbook for web application security requirements, security testing, and metrics. Use to establish security user stories and security use case release testing.   |
| OWASP Threat and Safeguard Matrix (TaSM)              | An action oriented view to safeguard and enable the business  | This matrix allows a company to overlay its major threats with the NIST Cyber Security Framework Functions (Identify, Protect, Detect, Respond, & Recover) to build a robust security plan. Use it as a dashboard to track and report on security across the organization. |
| Defect Dojo   | An open source vulnerability management tool that streamlines the testing process by offering templating, report generation, metrics, and baseline self-service tools.  | Use Defect Dojo to reduce the time for logging vulnerabilities with templates for vulnerabilities, imports for common vulnerability scanners, report generation, and metrics.  |

Table 4.1: OWASP Resources

**MITRE Resources** The increased frequency of LLM threats emphasizes the value of a resilience-first approach to defending an organization's attack surface. Existing TTPs are combined with new attack surfaces and capabilities in LLM Adversary threats and mitigations. MITRE maintains a well-established and widely accepted mechanism for coordinating opponent tactics and procedures based on real-world observations.

Coordination and mapping of an organization's LLM Security Strategy to MITRE ATT&CK and MITRE ATLAS allows an organization to determine where LLM Security is covered by current processes such as API Security Standards or where security holes exist.

MITRE ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge) is a framework, collection of data matrices, and assessment tool that was made by the MITRE Corporation to help organizations figure out how well their cybersecurity works across their entire digital attack surface and find holes that had not been found before. It is a knowledge repository that is used all over the world. The MITRE ATT&CK matrix contains a collection of strategies used by adversaries to achieve a certain goal. In the ATT&CK Matrix, these objectives are classified as tactics. The objectives are outlined in attack order, beginning with reconnaissance and progressing to the eventual goal of exfiltration or impact.

MITRE ATLAS, which stands for "Adversarial Threat Landscape for Artificial Intelligence Systems," is a knowledge base that is based on real-life examples of attacks on machine learning (ML) systems by bad actors. ATLAS is based on the MITRE ATT&CK architecture, and its tactics and procedures complement those found in ATT&CK.

| MITRE Resource        | Description   | Why It Is Recommended & Where To Use It  |
|-----------------------|---|--|
| MITRE ATT&CK          | Knowledge base of adversary tactics and techniques based on real-world observations | The ATT&CK knowledge base is used as a foundation for the development of specific threat models and methodologies. Map existing controls within the organization to adversary tactics and techniques to identify gaps or areas to test.      |
| MITRE AT&CK Workbench | Create or extend ATT&CK data in a local knowledge base                              | Host and manage a customized copy of the ATT&CK knowledge base. This local copy of the ATT&CK knowledge base can be extended with new or updated techniques, tactics, mitigation groups, and software that is specific to your organization. |

| MITRE Resource                         | Description  | Why It Is Recommended & Where To Use It   |
|--|--|---|
| MITRE ATLAS                            | MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) is a knowledge base of adversary tactics, techniques, and case studies for machine learning (ML) systems based on real-world observations, demonstrations from ML red teams and security groups, and the state of the possible from academic research | Use it to map known ML vulnerabilities and map checks and controls for proposed projects or existing systems.   |
| MITRE ATT&CK Powered Suit              | ATT&CK Powered Suit is a browser extension that puts the MITRE ATT&CK knowledge base at your fingertips.   | Add to your browser to quickly search for tactics, techniques, and more without disrupting your workflow.   |
| The Threat Report ATT&CK Mapper (TRAM) | Automates TTP Identification in CTI Reports  | Mapping TTPs found in CTI reports to MITRE ATT&CK is difficult, error prone, and time-consuming. TRAM uses LLMs to automate this process for the 50 most common techniques. Supports Jupyter notebooks. |
| Attack Flow v2.1.0                     | Attack Flow is a language for describing how cyber adversaries combine and sequence various offensive techniques to achieve their goals.   | Attack Flow helps visualize how an attacker uses a technique, so defenders and leaders understand how adversaries operate and improve their own defensive posture.                                      |
| MITRE Caldera                          | A cyber security platform (framework) designed to easily automate adversary emulation, assist manual red-teams, and automate incident response.  | Plugins are available for Caldera that help to expand the core capabilities of the framework and provide additional functionality, including agents, reporting, collections of TTPs and others          |
| CALDERA plugin: Arsenal                | A plugin developed for adversary emulation of AI-enabled systems.  | This plugin provides TTPs defined in MITRE ATLAS to interface with CALDERA.   |

| MITRE Resource       | Description  | Why It Is Recommended & Where To Use It   |
|----------------------|--|---|
| Atomic Red Team      | Library of tests mapped to the MITRE ATT&CK framework. | Use to validate and test controls in an environment. Security teams can use Atomic Red Team to quickly, portably, and reproducibly test their environments. You can execute atomic tests directly from the command line; no installation is required. |
| MITRE CTI Blueprints | Automates Cyber Threat Intelligence reporting.         | CTI Blueprints helps Cyber Threat Intelligence (CTI) analysts create high-quality, actionable reports more consistently and efficiently.  |

Table 4.2: MITRE Resources



## AI Vulnerability Repositories

| Name  | Description   |
|---|---|
| AI Incident Database  | A repository of articles about different times AI has failed in real-world applications and is maintained by a college research group and crowds sourced. |
| OECD AI Incidents Monitor (AIM)   | Offers an accessible starting point for comprehending the landscape of AI-related challenges.   |
| <b>Three of the leading companies tracking AI Model vulnerabilities</b> |   |
| Huntr Bug Bounty : ProtectAI  | Bug bounty platform for AI/ML   |
| AI Vulnerability Database (AVID) : Garak                                | Database of model vulnerabilities   |
| AI Risk Database: Robust Intelligence                                   | Database of model vulnerabilities   |

Table 4.3: AI Vulnerability Repositories

## AI Procurement Guidance

| Name  | Description   |
|---|---|
| World Economic Forum: Adopting AI Responsibly: Guidelines for Procurement of AI Solutions by the Private Sector: Insight Report June 2023 | <p>The standard benchmarks and assessment criteria for procuring Artificial systems are in early development. The procurement guidelines provide organizations with a baseline of considerations for the end-to-end procurement process.</p> <p>Use this guidance to augment an organization's existing Third Party Risk Supplier and Vendor procurement process.</p> |

Table 4.4: AI Procurement Guidance

# Team

Thank you to the OWASP Top 10 for LLM Applications Cybersecurity and Governance Checklist Contributors.

| Checklist Contributors    |                |                   |
|---------------------------|----------------|-------------------|
| Sandy Dunn                | Heather Linn   | John Sotiropoulos |
| Steve Wilson              | Fabrizio Cilli | Aubrey King       |
| Bob Simonoff              | David Rowe     | Rob Vanderveer    |
| Emmanuel Guilherme Junior | Andrea Succi   | Jason Ross        |
| Talesh Seeparsan          | Anthony Glynn  | Julie Tao         |

Table A.1: OWASP LLM AI Security & Governance Checklist  
Team

This project is licensed under the terms of the Creative Commons Attribution-ShareAlike 4.0 International License