**4.  Préparation d’une liste de *prompts* pour générer des offres de sécurité autour des modèles d’IA**

* Identifier les **risques de sécurité liés aux modèles IA**

(ex : prompt injection, data leakage, vol de modèle, hallucinations critiques, dérives éthiques…).

* **Injection de Prompt :** Permet aux attaquants de contourner le comportement du modèle, de faire fuiter des données confidentielles ou d'exécuter des instructions malveillantes en manipulant l'entrée.
* **Attaques par Inversion de Modèle (Model Inversion Attacks) :** Reconstruire des données d'entraînement sensibles à partir des sorties du modèle.
* **Empoisonnement de la Chaîne d'Approvisionnement (Modèle ou Jeu de Données) :** Introduire des données ou des composants malveillants dans le processus de développement ou de déploiement d'un modèle, altérant son comportement futur.
* **Abus d'API LLM & Attaques Basées sur le Débit :** Exploiter les API LLM pour des requêtes excessives, des attaques par déni de service, ou pour contourner les limites d'utilisation.
* **Jailbreaking via des Prompts Synthétiques :** Utiliser des prompts élaborés pour contourner les garde-fous du modèle et le faire générer du contenu inapproprié ou dangereux.
* **Data Leakage (Fuite de Données) :** Exposition involontaire de données sensibles à travers les sorties du modèle ou des mécanismes internes.
* **Vol de Modèle (Model Theft) :** Extraction illégale ou réplication d'un modèle propriétaire.
* **Hallucinations Critiques :** Génération par le modèle d'informations plausibles mais fausses, pouvant avoir des conséquences graves dans des contextes critiques.
* **Dérives Éthiques et Biais :** Reflet ou amplification des biais présents dans les données d'entraînement, conduisant à des résultats discriminatoires ou non éthiques.
* **Attaques par Contre-Exemples (Adversarial Attacks) :** Création d'entrées légèrement modifiées qui induisent en erreur le modèle, même si elles sont imperceptibles pour un humain.
* **Vulnérabilités dans les Bibliothèques et Frameworks IA :** Failles de sécurité au sein des logiciels et outils utilisés pour développer et déployer les modèles IA.
* **Manque de Transparence et d'Explicabilité (Explainability) :** Difficulté à comprendre pourquoi un modèle prend certaines décisions, rendant difficile l'identification et la correction des biais ou erreurs.

**Formulation d'une Série de Prompts Ciblés :**

**Objectif : Générer une offre de cybersécurité pour entreprise/organisation.**

* **Diagnostic :**
  + "Quels sont les principaux risques de sécurité liés à l'intégration de modèles d'IA générative (comme les LLM) dans une organisation du secteur [Nom du Secteur, ex : financier, santé] ? Fournissez une analyse SWOT des menaces potentielles."
  + "Élaborez un questionnaire d'évaluation des risques de cybersécurité spécifiques aux systèmes d'IA, couvrant l'ensemble du cycle de vie du modèle, de la conception au déploiement."
  + "Décrivez un scénario d'attaque par empoisonnement de données sur un modèle de détection de fraude basé sur l'IA et proposez des mesures d'atténuation techniques et organisationnelles."
* **Prévention :**
  + "Générez une proposition d'offre de services de cybersécurité pour la protection des infrastructures d'IA, incluant la sécurité des données d'entraînement, la résilience aux attaques par injection de prompt et la détection des fuites de modèles."
  + "Rédigez une politique de sécurité des LLM pour une entreprise, abordant les bonnes pratiques d'utilisation, la gestion des accès et la réponse aux incidents de sécurité liés aux modèles."
  + "Quelles solutions technologiques recommanderiez-vous pour protéger un système d'IA contre les attaques par inversion de modèle et les attaques par contre-exemples ?"
* **Conformité :**
  + "Comment une organisation peut-elle démontrer sa conformité aux réglementations comme le RGPD ou l'AI Act de l'UE en ce qui concerne la sécurité des données utilisées par ses modèles d'IA ?"
  + "Créez une liste de contrôles de sécurité (checklist) pour évaluer la conformité d'un système d'IA aux normes de sécurité cyber, en mettant l'accent sur la gouvernance et la traçabilité."

**Objectif : Décrire des cas d’usage de sécurité IA.**

* **Diagnostic :**
  + "Identifiez et décrivez au moins cinq cas d'usage où l'IA peut être utilisée pour renforcer la cybersécurité d'une entreprise (ex: détection d'anomalies, analyse de vulnérabilités)."
  + "Comment l'IA peut-elle aider à prévenir les attaques par phishing et l'ingénierie sociale en identifiant les schémas comportementaux suspects ?"
* **Prévention :**
  + "Expliquez comment les techniques d'apprentissage automatique peuvent être appliquées pour la détection précoce des menaces persistantes avancées (APT) dans un réseau d'entreprise."
  + "Décrivez un cas d'usage de l'IA pour l'automatisation de la réponse aux incidents de sécurité, en détaillant les étapes et les bénéfices attendus."
* **Sensibilisation :**
  + "Fournissez des exemples concrets de la manière dont les attaquants utilisent l'IA pour mener des cyberattaques sophistiquées, à des fins de sensibilisation des employés."

**Objectif : Simuler un plan d’audit de modèle IA.**

* **Audit :**
  + "Élaborez un plan d'audit détaillé pour évaluer la robustesse d'un modèle de classification d'images face aux attaques par contre-exemples."
  + "Quels sont les critères clés à inclure dans un audit de sécurité d'un LLM, en se concentrant sur la détection du jailbreaking et de l'injection de prompt ?"
  + "Décrivez les étapes d'un processus d'audit pour vérifier la non-divulgation de données sensibles par un modèle d'IA (fuite de données)."
  + "Proposez un cadre d'évaluation pour la résilience d'un modèle d'IA aux attaques par empoisonnement de la chaîne d'approvisionnement."
* **Diagnostic :**
  + "Quels outils et techniques d'évaluation sont essentiels pour un audit de sécurité complet des modèles d'IA, y compris l'analyse de la surface d'attaque et des vulnérabilités des API ?"

**Objectif : Proposer des actions de durcissement (hardening) ou de gouvernance autour des LLM.**

* **Prévention :**
  + "Proposez une liste de 10 actions de durcissement pour un LLM exposé via une API publique, incluant des mesures techniques et organisationnelles."
  + "Décrivez les meilleures pratiques pour la gestion des données sensibles utilisées pour l'entraînement des LLM, afin de minimiser le risque de fuite de données."
  + "Comment peut-on mettre en œuvre des mécanismes de filtrage d'entrée et de sortie robustes pour un LLM afin de contrer les tentatives d'injection de prompt et de jailbreaking ?"
  + "Quelles stratégies de sécurisation des API devraient être mises en place pour protéger un LLM contre les abus d'API et les attaques basées sur le débit ?"
* **Gouvernance :**
  + "Élaborez un cadre de gouvernance pour l'utilisation éthique et sécurisée des LLM au sein d'une organisation, couvrant les responsabilités, les processus de validation et les mécanismes de supervision."
  + "Comment établir un programme de Bug Bounty spécifique aux LLM pour identifier et corriger les vulnérabilités liées à l'injection de prompt et aux hallucinations critiques ?"
  + "Quelles sont les responsabilités clés d'une équipe de sécurité dans la gestion du cycle de vie sécurisé des modèles d'IA ?"
* **Sensibilisation :**
  + "Rédigez un guide de bonnes pratiques pour les développeurs et les utilisateurs de LLM, axé sur la sécurité et la prévention des dérives."

**Classification des Prompts par Objectif :**

Pour une meilleure organisation, les prompts sont classés par objectif principal, bien que certains puissent relever de plusieurs catégories.

**1. Diagnostic :**

* "Quels sont les principaux risques de sécurité liés à l'intégration de modèles d'IA générative (comme les LLM) dans une organisation du secteur [Nom du Secteur, ex : financier, santé] ? Fournissez une analyse SWOT des menaces potentielles."
* "Élaborez un questionnaire d'évaluation des risques de cybersécurité spécifiques aux systèmes d'IA, couvrant l'ensemble du cycle de vie du modèle, de la conception au déploiement."
* "Décrivez un scénario d'attaque par empoisonnement de données sur un modèle de détection de fraude basé sur l'IA et proposez des mesures d'atténuation techniques et organisationnelles."
* "Identifiez et décrivez au moins cinq cas d'usage où l'IA peut être utilisée pour renforcer la cybersécurité d'une entreprise (ex: détection d'anomalies, analyse de vulnérabilités)."
* "Comment l'IA peut-elle aider à prévenir les attaques par phishing et l'ingénierie sociale en identifiant les schémas comportementaux suspects ?"
* "Quels outils et techniques d'évaluation sont essentiels pour un audit de sécurité complet des modèles d'IA, y compris l'analyse de la surface d'attaque et des vulnérabilités des API ?"

**2. Prévention :**

* "Générez une proposition d'offre de services de cybersécurité pour la protection des infrastructures d'IA, incluant la sécurité des données d'entraînement, la résilience aux attaques par injection de prompt et la détection des fuites de modèles."
* "Rédigez une politique de sécurité des LLM pour une entreprise, abordant les bonnes pratiques d'utilisation, la gestion des accès et la réponse aux incidents de sécurité liés aux modèles."
* "Quelles solutions technologiques recommanderiez-vous pour protéger un système d'IA contre les attaques par inversion de modèle et les attaques par contre-exemples ?"
* "Expliquez comment les techniques d'apprentissage automatique peuvent être appliquées pour la détection précoce des menaces persistantes avancées (APT) dans un réseau d'entreprise."
* "Décrivez un cas d'usage de l'IA pour l'automatisation de la réponse aux incidents de sécurité, en détaillant les étapes et les bénéfices attendus."
* "Proposez une liste de 10 actions de durcissement pour un LLM exposé via une API publique, incluant des mesures techniques et organisationnelles."
* "Décrivez les meilleures pratiques pour la gestion des données sensibles utilisées pour l'entraînement des LLM, afin de minimiser le risque de fuite de données."
* "Comment peut-on mettre en œuvre des mécanismes de filtrage d'entrée et de sortie robustes pour un LLM afin de contrer les tentatives d'injection de prompt et de jailbreaking ?"
* "Quelles stratégies de sécurisation des API devraient être mises en place pour protéger un LLM contre les abus d'API et les attaques basées sur le débit ?"

**3. Conformité :**

* "Comment une organisation peut-elle démontrer sa conformité aux réglementations comme le RGPD ou l'AI Act de l'UE en ce qui concerne la sécurité des données utilisées par ses modèles d'IA ?"
* "Créez une liste de contrôles de sécurité (checklist) pour évaluer la conformité d'un système d'IA aux normes de sécurité cyber, en mettant l'accent sur la gouvernance et la traçabilité."

**4. Audit :**

* "Élaborez un plan d'audit détaillé pour évaluer la robustesse d'un modèle de classification d'images face aux attaques par contre-exemples."
* "Quels sont les critères clés à inclure dans un audit de sécurité d'un LLM, en se concentrant sur la détection du jailbreaking et de l'injection de prompt ?"
* "Décrivez les étapes d'un processus d'audit pour vérifier la non-divulgation de données sensibles par un modèle d'IA (fuite de données)."
* "Proposez un cadre d'évaluation pour la résilience d'un modèle d'IA aux attaques par empoisonnement de la chaîne d'approvisionnement."

**5. Gouvernance :**

* "Élaborez un cadre de gouvernance pour l'utilisation éthique et sécurisée des LLM au sein d'une organisation, couvrant les responsabilités, les processus de validation et les mécanismes de supervision."
* "Comment établir un programme de Bug Bounty spécifique aux LLM pour identifier et corriger les vulnérabilités liées à l'injection de prompt et aux hallucinations critiques ?"
* "Quelles sont les responsabilités clés d'une équipe de sécurité dans la gestion du cycle de vie sécurisé des modèles d'IA ?"

**6. Sensibilisation :**

* "Fournissez des exemples concrets de la manière dont les attaquants utilisent l'IA pour mener des cyberattaques sophistiquées, à des fins de sensibilisation des employés."
* "Rédigez un guide de bonnes pratiques pour les développeurs et les utilisateurs de LLM, axé sur la sécurité et la prévention des dérives."