Sécurité des LLM en Production : Injection de Prompt, Exfiltration de Données et Attaques de Jailbreak

Lien : https://innov-systems.com/formation/securite-des-llm-en-production-injection-de-prompt-exfiltration-de-donnees-et-attaques-de-jailbreak

DURÉE
5 jours (35h)

RÉFÉRENCE
BDI160

CATÉGORIE
Sécurité de l'Intelligence Artificielle

Innov Systems

OBJECTIFS DE LA FORMATION

À l'issue de cette formation, vous serez capable de :

- Comprendre les risques de sécurité liés aux LLM
- Identifier les attaques de prompt injection et jailbreak
- Prévenir l’exfiltration de données sensibles
- Sécuriser les pipelines d’IA en production
- Mettre en place des mécanismes de défense adaptés
- Surveiller les comportements anormaux des modèles
- Appliquer les bonnes pratiques de prompt engineering sécurisé
- Réduire la surface d’attaque des systèmes LLM
- Intégrer la sécurité dans le cycle de vie des modèles
- Réaliser des audits de sécurité sur systèmes IA
- Comprendre les enjeux de conformité et gouvernance
- Étudier des cas réels d’attaques sur LLM

POUR QUI ?

- Ingénieurs IA
- Machine learning engineers
- Data scientists
- Architectes IA
- DevSecOps engineers
- Cybersecurity engineers
- Product managers IA
- Consultants en cybersécurité
- Responsables IT
- Chefs de projet IA
- Researchers en sécurité IA
- Toute personne impliquée dans le déploiement de LLM en production

Innov Systems

Programme détaillé

1 / Introduction à la sécurité des LLM en production

- Enjeux de sécurité des modèles de langage en entreprise
- Surface d’attaque des systèmes LLM
- Risques liés à l’usage en production

2 / Comprendre l’architecture des LLM déployés

- Chaîne d’inférence et orchestration
- Intégration avec APIs et outils externes
- Points de vulnérabilité typiques

3 / Prompt injection : principes et mécanismes

- Définition de l’injection de prompt
- Attaques directes et indirectes
- Contournement des instructions système

4 / Attaques de jailbreak sur les LLM

- Techniques de contournement des garde-fous
- Manipulation du contexte conversationnel
- Évolution des méthodes d’attaque

5 / Exfiltration de données sensibles

- Risques de fuite via prompts malveillants
- Accès indirect aux données internes
- Scénarios d’attaque en entreprise

Innov Systems

6 / Sécurisation des chaînes LLM (LLM pipelines)

- Filtrage des entrées et sorties
- Isolation des données sensibles
- Gestion des permissions et outils

7 / Défenses techniques contre les attaques

- Prompt hardening et system prompts robustes
- Modération automatique et classifiers
- Red teaming des modèles

8 / Monitoring et détection des comportements malveillants

- Logging des interactions LLM
- Détection d’anomalies et signaux faibles
- Audit de sécurité des prompts

9 / Gouvernance et conformité des LLM

- Politiques d’usage des modèles
- Conformité RGPD et protection des données
- Responsabilités des équipes IA

10 / Études de cas et exercices pratiques

- Analyse d’attaques réelles sur LLM
- Scénarios de jailbreak contrôlés
- Bonnes pratiques de sécurisation en production

Innov Systems

Approche pédagogique

- Support Ecrit et Projection
- Exposés Interactifs, Podcasts et Vidéos
- Brainstorming et Jeux de Rôle
- Cas Pratiques et Labs inclus pour leur impact opérationnel
- Test de Validation des Acquis des Connaissances

Prochaines dates programmées

15 au 19 Fév. 2027	Présentiel - Casablanca
12 au 16 Avr. 2027	Présentiel - Casablanca
07 au 11 Juin 2027	Présentiel - Casablanca
02 au 06 Août 2027	Présentiel - Casablanca
27 Sep. au 01 Oct. 2027	Présentiel - Casablanca

Autres dates possibles sur demande. Contactez-nous pour organiser une session intra-entreprise.

Réservation & Renseignements

Téléphone : +212 522 247 210
Email : contact@innov-systems.com
Web : https//www.innov-systems.com

Innov Systems — Tous droits réservés