AI Red Teaming : Méthodologies d'Attaque et de Test Adversarial des Modèles d'IA

Lien : https://innov-systems.com/formation/ai-red-teaming-methodologies-dattaque-et-de-test-adversarial-des-modeles-dia

DURÉE
5 jours (35h)

RÉFÉRENCE
BDI154

CATÉGORIE
Sécurité de l'Intelligence Artificielle

Innov Systems

OBJECTIFS DE LA FORMATION

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

- Comprendre les principes du AI Red Teaming
- Identifier les vulnérabilités des modèles IA
- Maîtriser les techniques d’attaque adversariale
- Tester la robustesse des LLM et agents IA
- Concevoir des scénarios d’attaque réalistes
- Évaluer la sécurité des systèmes IA
- Mettre en place des méthodologies de test structurées
- Améliorer la résilience des modèles IA
- Intégrer le red teaming dans les cycles MLOps
- Utiliser des outils d’audit de sécurité IA
- Développer une culture de sécurité IA
- Étudier des cas réels de failles IA

POUR QUI ?

- Ingénieurs cybersécurité
- AI security engineers
- Data scientists
- Machine learning engineers
- Pentesters
- DevSecOps engineers
- Architectes IA
- Consultants en cybersécurité
- Chefs de projet IA
- Responsables sécurité des systèmes d’information
- Chercheurs en IA
- Toute personne impliquée dans la sécurité des modèles IA

Innov Systems

Programme détaillé

1 / Introduction au AI Red Teaming

- Définition du red teaming appliqué à l'IA
- Objectifs des tests adversariaux
- Enjeux de sécurité des modèles IA

2 / Menaces et surfaces d'attaque des systèmes IA

- Prompt injection et jailbreaks
- Data poisoning et manipulation des données
- Fuites d'informations et hallucinations exploitées

3 / Fondamentaux de l'attaque adversariale

- Notions d'exemples adversariaux
- Vulnérabilités des modèles de machine learning
- Comportements non robustes des LLM

4 / Méthodologies de test red team

- Approches systématiques de test
- Scénarios d'attaque structurés
- Simulation d'utilisateurs malveillants

5 / Tests sur modèles LLM et agents IA

- Attaques sur chaînes de prompts
- Exploitation des outils et plugins
- Failles dans les agents autonomes

Innov Systems

6 / Évaluation de la robustesse des modèles

- Métriques de sécurité IA
- Taux de succès des attaques
- Analyse des points de défaillance

7 / Défenses et stratégies de mitigation

- Guardrails et filtres de sécurité
- Alignement et fine-tuning défensif
- Monitoring en production

8 / Red teaming en environnement entreprise

- Organisation d’une équipe red team IA
- Cycles de test continu
- Intégration dans le MLOps

9 / Outils et frameworks de sécurité IA

- Frameworks de test adversarial
- Outils d’audit de modèles
- Simulation d’attaques automatisées

10 / Études de cas et exercices pratiques

- Tests sur modèles réels
- Analyse de vulnérabilités connues
- Mise en place d’un exercice red team complet

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

08 au 12 Fév. 2027	Présentiel - Casablanca
05 au 09 Avr. 2027	Présentiel - Casablanca
31 Mai au 04 Juin 2027	Présentiel - Casablanca
26 au 30 Juil. 2027	Présentiel - Casablanca
20 au 24 Sep. 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