REDGUARD
SECURING YOUR ASSETS

Sicherheitsüberprüfung für Ihre KI-Systeme

Unzureichend gesicherte KI-Systeme können unabsichtlich vertrauliche Informationen preisgeben, unvorhergesehene Aktionen auslösen oder schädliche Antworten geben. Das hat mögliche Folgen für Ihre Assets und Ihr Image sowie eventuell auch rechtliche Konsequenzen.

WAS IST RED TEAMING FÜR GENAI?

Speziell entwickelte Angriffe prüfen Ihre generative KI auf ihre Anfälligkeit gegenüber verschiedenen Bedrohungen. Ziel ist es, Risiken aufzudecken und zu kombinieren, um relevante Angriffspfade zu identifizieren und so Sicherheit und Resilienz nachhaltig zu verbessern.

Jailbreak-Versuche

Umgehen der vordefinierten Sicherheitsmechanismen, um das Modell zu unerwünschten Handlungen zu bewegen

Datenextraktion

Das System wird dazu gebracht, sensible Informationen aus seinen Trainingsdaten preiszugeben

Halluzination & Bias

Aufdecken von Fehlinformationen, erfundenen Fakten und Vorurteilen in den Antworten

Prompt-Injection

Direkte und indirekte Manipulation der Eingabe, um das System zu verwirren und unerwünschte Aktionen auszulösen

Lassen Sie Ihre KI-Systeme auf ihre Sicherheit überprüfen!



Patrick Schmid
Partner & Head of Security Testing
Tel. +41 78 614 99 71
patrick.schmid@redguard.ch

REDGUARD
SECURING YOUR ASSETS

Risiken von KI-Systemen & Test-Vorgehen

WIESO IHRE KI-SYSTEME ÜBERPRÜFEN LASSEN?

KI-Modelle basieren auf komplexen, breit trainierten Grundlagen und wissen automatisch mehr, als man zunächst denkt:

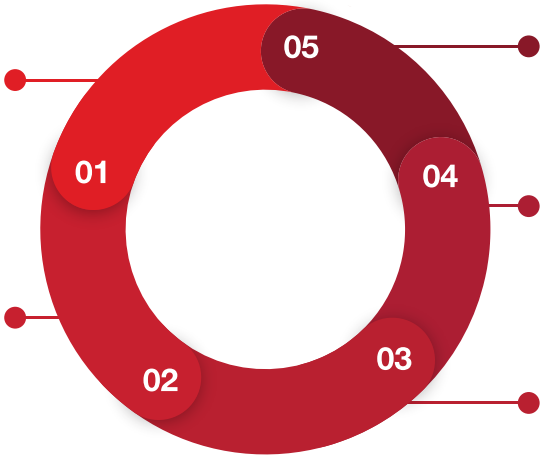
- **Quellenbeschränkung:** Cleveres Prompting kann die Systeminstruktionen der KI umgehen und Antworten aus dem breiten KI-Basiswissen freischalten.
- **Generative Natur der Modelle:** Antworten werden neu erzeugt und können irrelevante, irreführende oder falsche Schlüsse enthalten.
- **Implizites Wissen & Assoziationen:** Modelle übernehmen subtile Muster, Meinungen, Stereotype und Vorurteile aus den Trainingsdaten.
- **Prompt Injection & Jailbreaking:** Böswillige Prompts können das Modell dazu bringen, Anweisungen zu ignorieren und schädliche oder sensible Informationen auszugeben.

UNSER VORGEHEN NACH OWASP

Zur strukturierten Prüfung orientieren wir uns am OWASP GenAI Red Teaming Guide:

Kick-off und Initialisierung
Einrichten von relevanten Zugängen und Zugriffen sowie zur Verfügung stellen von Dokumenten und Informationen

Threat Modeling
Definieren und Priorisieren von zu prüfenden Angriffsszenarien im Zusammenhang mit der KI



Berichterstattung
Zusammenfassen der identifizierten Risiken, Schwachstellen und Schwächen in einem detaillierten Report

Statistical Assessment
Tool-basiertes, statistisches Prüfen der Bereiche Prompt Testing, Content Filtering und Adversarial Queries

Manual Application Stack Assessment
Prüfen der Teilbereiche Model Evaluation, Implementation Checks & Guardrails, System Testing und Runtime Interactions

Lassen Sie Ihre KI-Systeme auf ihre Sicherheit überprüfen!



Patrick Schmid
Partner & Head of Security Testing
Tel. +41 78 614 99 71
patrick.schmid@redguard.ch