Security & Guardrails Assessment for GenAI

Assessment | 1 Tag | Remote / Vor Ort für sensible Systeme möglich

GenAI-Systeme führen völlig neue Angriffsvektoren ein – von Prompt-Injection bis hin zu Datenexfiltration – die mit herkömmlichen Sicherheitsmaßnahmen nicht bekämpft werden können. Diese Bewertung verwandelt die Unsicherheit hinsichtlich der GenAI-Sicherheit in eine konkrete, umsetzbare Verteidigungsstrategie.

Unternehmen, die GenAI einsetzen, stehen vor einzigartigen Sicherheitsherausforderungen, die weit über die herkömmliche Anwendungssicherheit hinausgehen. Prompt-Injection, Jailbreaking, Datenlecks durch Modellausgaben und Missbrauchsszenarien erfordern spezielle Bedrohungsmodelle und Schutzstrategien. Diese gezielte Bewertung bringt Ihre Sicherheits- und Plattformteams mit unseren KI-Sicherheitsexperten zusammen, um systematisch GenAI-spezifische Bedrohungen zu identifizieren, ein umfassendes Bedrohungsmodell zu entwickeln und wirksame Schutzmaßnahmen und Kontrollen zu definieren. Sie erhalten eine priorisierte Sicherheits-Roadmap, konkrete Testfälle zur Validierung und Implementierungshinweise – so werden GenAI-Sicherheitsrisiken zu einer überschaubaren, kontrollierten Bereitstellung.

Setup

Zielunternehmen	Große Unternehmen, KMUs, öffentliche Organisationen
Reifegrad	Experimenter, Practitioner, Professional
Teilnehmer	Security Engineers, AppSec Specialists, Platform Engineers
iteratec	AI Security Engineer, AppSec Specialist
Voraussetzungen	Zugriff auf Architektur-Dokumentation und Testumgebung

Agenda

1. **Analyse der Bedrohungslage**
   Systematische Identifizierung von GenAI-spezifischen Angriffsvektoren
2. **Entwicklung eines Bedrohungsmodells**
   Erstellung eines detaillierten Bedrohungsmodells mit Risikobewertung (Wahrscheinlichkeit × Auswirkung) für Ihre spezifische GenAI-Anwendung
3. **Entwurf von Kontrollmechanismen**
   Definition von Strategien zur Eingabevalidierung, Mechanismen zur Ausgabefilterung und Mustern für die Zugriffskontrolle
4. **Rahmenwerk für Sicherheitstests**
   Entwicklung von Testfällen einschließlich gegnerischer Beispiele zur Validierung
5. **Überwachung und Reaktionsplanung**
   Entwurf von Verfahren für Sicherheitsprotokollierung, Anomalieerkennung und Reaktion auf Vorfälle

Erfolge

- **Umfassendes Bedrohungsmodell**
  Klares Verständnis der spezifischen Sicherheitsrisiken von GenAI mit quantifizierter Wahrscheinlichkeit und geschäftlichen Auswirkungen für Ihre Anwendung
- **Priorisierte Sicherheit**
  Roadmap Umsetzbarer Kontrollplan mit konkreten Leitplanken, geordnet nach Risikominderung und Implementierungsaufwand
- **Validierte Verteidigungsstrategie**
  Getestete Sicherheitsmaßnahmen mit spezifischen Testfällen zur Verhinderung von Prompt-Injection, Datenexfiltration und Systemmissbrauch

Ergebnisse

- ✓ **Detailliertes Bedrohungsmodell** mit Risikobewertungsmatrix und Dokumentation der Angriffsvektoren
- ✓ **Priorisierter Kontrollplan** mit Implementierungsleitfaden für Schutzvorkehrungen und Sicherheitsmaßnahmen
- ✓ **Security Test Suite** mit gegensätzlichen Beispielen und Validierungsszenarien
- ✓ **Monitoring Concept** mit Protokollierungsanforderungen und Anomalieerkennungsmustern
- ✓ **Incident Response Playbook** für GenAI-spezifische Sicherheitsvorfälle

Security & Guardrails Assessment for GenAI

Assessment | 1 day | Remote / On-site possible for sensitive systems

GenAI systems introduce entirely new attack vectors – from prompt injection to data exfiltration – that traditional security measures don't address. This assessment transforms GenAI security uncertainty into a concrete, actionable defense strategy.

Organizations deploying GenAI face unique security challenges that go far beyond conventional application security. Prompt injection, jailbreaking, data leakage through model outputs, and abuse scenarios require specialized threat modeling and protection strategies. This focused assessment brings together your security and platform teams with our AI security experts to systematically identify GenAI-specific threats, develop a comprehensive threat model, and define effective guardrails and controls. You leave with a prioritized security roadmap, concrete test cases for validation, and implementation guidance – turning GenAI security risks into manageable, controlled deployment.

Setup

Target Companies	Large Corporations, SMBs, Public Organisations
Maturity Level	Experimenter, Practitioner, Professional
Participants	Security Engineers, AppSec Specialists, Platform Engineers
iteratec	AI Security Engineer, AppSec Specialist
Prerequisites	Access to architecture documentation and test environment

Agenda

1. **Threat Landscape Analysis**
   Systematic identification of GenAI-specific attack vectors
2. **Threat Model Development**
   Creation of detailed threat model with risk assessment (likelihood × impact) for your specific GenAI application
3. **Control Design**
   Definition of input validation strategies, output filtering mechanisms, and access control patterns
4. **Security Testing Framework**
   Development of test cases including adversarial examples for validation
5. **Monitoring & Response Planning**
   Design of security logging, anomaly detection, and incident response procedures

Achievements

- **Comprehensive Threat Model**
  Clear understanding of GenAI-specific security risks with quantified likelihood and business impact for your application
- **Prioritized Security**
  Roadmap Actionable control plan with concrete guardrails, ranked by risk reduction and implementation effort
- **Validated Defense Strategy**
  Tested security measures with specific test cases to prevent prompt injection, data exfiltration, and system abuse

Deliverables

- ✓ **Detailed Threat Model** with risk assessment matrix and attack vector documentation
- ✓ **Prioritized Control Plan** with implementation guidance for guardrails and security measures
- ✓ **Security Test Suite** with adversarial examples and validation scenarios
- ✓ **Monitoring Concept** with logging requirements and anomaly detection patterns
- ✓ **Incident Response Playbook** for GenAI-specific security incidents