

# Red-Team Day

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

**Ihr GenAI-System mag produktionsreif erscheinen – bis jemand versucht, es zu knacken. In dieser intensiven Red-Team-Sitzung wird Ihr System systematisch angegriffen, um Schwachstellen, Halluzinationen und Qualitätsprobleme aufzudecken, bevor Ihre Benutzer dies tun.**

GenAI-Systeme können auf unerwartete Weise versagen, die bei herkömmlichen Tests nicht erkannt wird – von subtilen Halluzinationen bis hin zu katastrophalen Prompt-Injektionen. An diesem praxisorientierten Red-Team-Tag kommen Ihre Produkt- und Engineering-Teams mit unseren KI-Sicherheitsexperten zusammen, um intensive Adversarial-Tests durchzuführen. Wir greifen Ihr System systematisch mit realistischen Techniken an, identifizieren Schwachstellen in Ihren Sicherheitsvorkehrungen und decken Qualitätsprobleme auf, die das Vertrauen der Benutzer untergraben könnten. Sie erhalten einen detaillierten Bericht mit den Ergebnissen, priorisierten Korrekturen und einer umfassenden Testsuite für die kontinuierliche Qualitätssicherung – so werden Sicherheitsbedenken in messbare Verbesserungen umgewandelt.

## Setup

<b>Zielunternehmen</b>	Große Unternehmen, KMUs, öffentliche Organisationen
<b>Reifegrad</b>	Experimenter, Practitioner, Professional
<b>Teilnehmer</b>	Product/Platform Teams, Architects, AI Engineers, QA Leads
<b>iteratec</b>	AI Security Engineer
<b>Voraussetzungen</b>	Zugriff auf Architektur-Dokumentation und Testumgebung

## Agenda

- 1. Attack Surface Mapping**  
Systematische Identifizierung potenzieller Schwachstellen und Angriffsvektoren in Ihrer GenAI-Anwendung
- 2. Adversarial Testing**  
Praktische Angriffe, darunter Prompt-Injection, Jailbreaking und Versuche der Kontextmanipulation
- 3. Guardrail Validation**  
Stresstests für implementierte Sicherheitsmaßnahmen und Content-Moderationssysteme
- 4. Edge Case Discovery**  
Identifizierung problematischer Randbedingungen und Ausfallmodi
- 5. Test Suite Development**  
Erstellung wiederverwendbarer gegnerischer Testfälle für Regressionstests

## Erfolge

- **Umfassender Schwachstellenbericht**  
Detaillierte Ergebnisse mit Schweregradklassifizierung und konkreten Beispielen für erfolgreiche Angriffe und Qualitätsprobleme
- **Umsetzbarer Lösungsplan**  
Priorisierte Empfehlungen mit Aufwandsschätzungen, um identifizierte Schwachstellen systematisch zu beheben
- **Produktionsreife Testsuite**  
50–100 gegnerische Testfälle, bereit für die CI/CD-Integration, um Regressionen zu verhindern

## Ergebnisse

- ✓ **Detaillierter Ergebnisbericht** mit Schweregradklassifizierung und Angriffsdokumentation
- ✓ **Priorisierte Korrekturvorschläge** mit Schätzungen zum Implementierungsaufwand
- ✓ **Empfehlungen zur kontinuierlichen Bewertung** für eine fortlaufende Qualitätssicherung

# Red-Team Day

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

**Your GenAI system might look production-ready – until someone tries to break it. This intensive red-team session systematically attacks your system to uncover vulnerabilities, hallucinations, and quality issues before your users do.**

GenAI systems can fail in unexpected ways that traditional testing doesn't catch – from subtle hallucinations to catastrophic prompt injections. This hands-on red-team day brings together your product and engineering teams with our AI security experts for intensive adversarial testing. We systematically attack your system using real-world techniques, identify weaknesses in your guardrails, and uncover quality issues that could undermine user trust. You leave with a detailed findings report, prioritized fixes, and a comprehensive test suite for ongoing quality assurance – transforming security concerns into measurable improvements.

## Setup

<b>Target Companies</b>	Large Corporations, SMBs, Public Organisations
<b>Maturity Level</b>	Experimenter, Practitioner, Professional
<b>Participants</b>	Product/Platform Teams, Architects, AI Engineers, QA Leads
<b>iteratec</b>	AI Security Engineer
<b>Prerequisites</b>	Access to architecture documentation and test environment

## Agenda

- 1. Attack Surface Mapping**  
Systematic identification of potential vulnerability points and attack vectors in your GenAI application
- 2. Adversarial Testing**  
Hands-on attacks including prompt injection, jailbreaking, and context manipulation attempts
- 3. Guardrail Validation**  
Stress-testing implemented safety measures and content moderation systems
- 4. Edge Case Discovery**  
Identification of problematic boundary conditions and failure modes
- 5. Test Suite Development**  
Creation of reusable adversarial test cases for regression testing

## Achievements

- **Comprehensive Vulnerability Report**  
Detailed findings with severity classification and concrete examples of successful attacks and quality issues
- **Actionable Fix Roadmap**  
Prioritized recommendations with effort estimates to address identified weaknesses systematically
- **Production-Ready Test Suite**  
50-100 adversarial test cases ready for CI/CD integration to prevent regression

## Deliverables

- ✓ **Detailed Findings Report** with severity classification and attack documentation
- ✓ **Prioritized Fix Recommendations** with implementation effort estimates
- ✓ **Continuous Evaluation Recommendations** for ongoing quality assurance