PRAETORIAN

AI Red Teaming

Emulate real-world attacks against AI systems

salesforce TOYOTA stripe Abbott NETFLIX

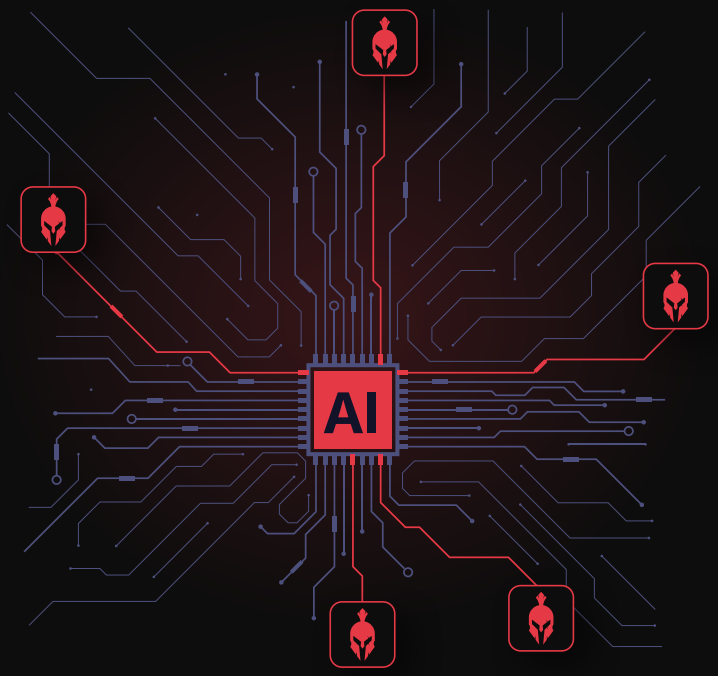
Solution Overview

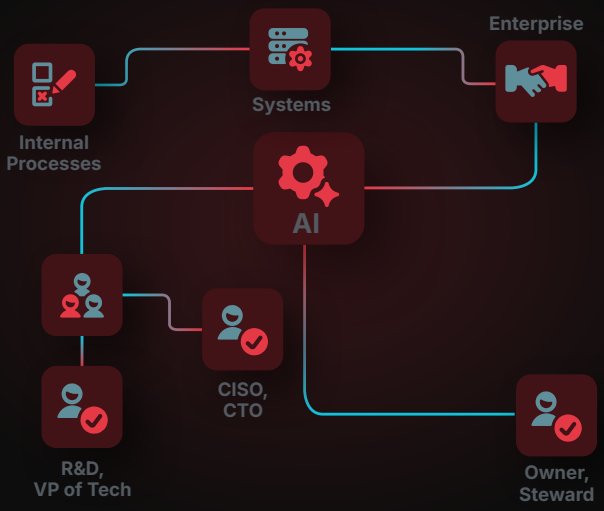
Artificial intelligence introduces new classes of vulnerabilities that traditional security testing fails to uncover. Risks that can silently undermine enterprise security, brand integrity, and innovation investments.

Praetorian's AI Red Teaming service applies our industry-leading offensive security expertise to your organization's GenAI systems to simulate real-world attacks, assess defensive readiness, and deliver actionable recommendations for improvement.

Our engagements are designed to identify meaningful security gaps, avoid distraction from overhyped or low-impact issues, and demonstrate real-world exploitation scenarios that matter most to your business and R&D investment. This results in a technically rigorous assessment that delivers clear, prioritized insights.

Praetorian operators leverage advanced adversarial techniques such as RAG database poisoning, model theft, indirect prompt injection, and excessive agency to expose risk and measure true impact.





Who This Is For

- ✓ Enterprises embedding GenAI technology into products, systems, or internal processes
- ✓ Owners and stewards of organizational AI or GenAI risk
- ✓ CISOs, CTOs, Product Development VPs, and R&D Directors responsible for securing AI initiatives

Copyright Praetorian Security. All Rights Reserved.

PRAETORIAN

Service Offerings

Praetorian offers three complementary engagement types, tailored to the maturity, complexity, and objectives of your GenAI program.

GenAI Penetration Test

Guided by: OWASP Top 10 for LLM Applications

Purpose: Designed for organizations integrating GenAI functionality at the application layer and beginning to formalize AI security testing.

Consists of:

- ✓ A thorough review of the most salient LLM-application vulnerabilities
- ✓ Demonstration of real-world exploitability and business impact

Most beneficial for:

- ✓ Developers of LLM-enabled applications
- ✓ Organizations already familiar with other OWASP Top 10 frameworks

Deliverables:

- ✓ Executive summary
- ✓ Vulnerability listing with CVSS 4.0 ratings
- ✓ Proof of exploitation
- ✓ Step-by-step reproduction guidance
- ✓ Recommendations for improvement

Deliverables:
2-4 weeks

GenAI Attack Path Mapping

Guided by: MITRE ATLAS™

Purpose: For organizations with more complex GenAI environments seeking to understand how vulnerabilities chain together to produce material risk and drive change.

Consists of:

- ✓ Exploitation of multi-stage vulnerability chains across GenAI systems
- ✓ Targeting of specific objectives defined by the client to illustrate business impact

Most beneficial for:

- ✓ Developers of complex GenAI systems
- ✓ Organizations using GenAI beyond LLMs alone
- ✓ Teams familiar with *MITRE ATT&CK®*

Deliverables:

All items from the GenAI Penetration Test, plus:

- ✓ Detailed attack-chain diagram

Deliverables:
4-8 weeks

GenAI Red Team Operation

Focus: Stealth and Evasion

Purpose: A capstone engagement for organizations with significant AI investment—such as frontier-model developers or enterprises operating advanced GenAI infrastructure—seeking to evaluate detection and response under realistic adversarial conditions.

Consists of:

- ✓ Stealth & evasion red team operation targeting the organization's most sensitive AI assets
- ✓ Targeting of the underlying AI infrastructure to steal sensitive models, proprietary information, and R&D investments
- ✓ Optional follow-on purple team exercise for defensive tuning

Most beneficial for:

- ✓ Frontier model developers
- ✓ Orgs deploying complex GenAI-related infrastructure
- ✓ Orgs operating extensive GPU compute capabilities

Deliverables:

All items from the GenAI Attack Path Mapping, plus:

- ✓ Comprehensive attack narrative
- ✓ Optional Purple Team report

Deliverables:
8-12 weeks

MITRE ATLAS™ and MITRE ATT&CK® are a trademark and registered trademark of The MITRE Corporation.

Copyright Praetorian Security. All Rights Reserved.