OWASP | GenAI SECURITY PROJECT
TOP 10 FOR LLM AND GENERATIVE AI

# Vendor Evaluation Criteria for AI Red Teaming Providers & Tooling

*Clear, practical criteria for evaluating vendors of AI Red Teaming consulting services and automated tools, across simple and advanced GenAI systems*

Version 1.0 – Public Release
January 13, 2026

The information provided in this document does not, and is not intended to, constitute legal advice. All information is for general informational purposes only. This document contains links to other third-party websites. Such links are only for convenience and OWASP does not recommend or endorse the contents of the third-party sites.

## License and Usage

# Table of Content

# 1. Executive Summary

Most organizations deploy simple GenAI systems, such as chatbots, customer-support assistants, workflow copilots, and basic RAG applications. A smaller but growing number operate advanced AI systems such as tool-calling agents, MCP-based architectures, and multi-agent workflows.

Both categories introduce risks that require specialized adversarial evaluation. Many vendors overclaim, focusing on superficial jailbreak demonstrations or static prompt libraries while ignoring systemic vulnerabilities, workflow bypasses, misuse risks, and failures in agentic or tool-enabled systems.

This document provides criteria for evaluating vendors of AI Red Teaming consulting services as well as automated tooling, and includes considerations for both simple and advanced GenAI deployments. It highlights:

- What effective AI red teaming looks like
- What questions to ask vendors
- What "green flags" indicate competence
- What "red flags" signal low-quality or misleading offerings
- How to distinguish genuine adversarial evaluation from superficial testing

The goal is to enable business leaders and executives to confidently identify providers who reduce real risk, not those who merely offer the illusion of coverage.

## 1.1. Quick Executive Guide: Green Flags & Red Flags

Here are some very high-level items you can use as a guide to quickly gauge the level of competence of a vendor you are exploring a partnership with.

### ✅ Green Flags

- Reproducible single-turn and multi-turn adversarial evaluations
- Custom testing with novel findings, not recycled jailbreaks from public databases
- Clear reporting with business impact mapping
- Demonstrated capability and experience across both simple and advanced systems
- Human verification of critical findings
- Ability to evaluate stateful systems, including memory, long-lived sessions, and cross-session behavior

- Findings translated into actionable remediation guidance

## 🚫 Red Flags

- Stock jailbreak libraries passed off as red teaming
- Vendor cannot articulate how modern architectures (including RAG, Memory, etc), protocols (MCP, A2A, ACP), and methods (fine-tuning, tool calling) work
- No multi-turn or stateful test capability
- AI-generated evaluations with no human oversight
- Claims of "full coverage" or one-click red teaming
- Focus exclusively on model outputs, ignoring actions, system state changes, or real-world impact
- Lack of functional testing
- Black-box scoring with no transparency into methodology or test design
- Absence of use-case-specific and customized evaluations, with no adaptation to the customer's real workflows, data, or business logic

# 2. Background and Definitions

## 2.1 What AI Red Teaming Is (and is not)

For the purpose of this document, AI Red Teaming is defined as the adversarial testing of AI systems to uncover safety, security, misuse, robustness, ethical, and alignment failure modes. It includes both single-turn and multi-turn probing, workflow abuse, systemic bypasses, and scenario-based misuse discovery.

This document does **not** address vendors who use AI to perform traditional red teaming of infrastructure, networks, web apps, or social engineering. It is strictly focused on vendors providing services to perform red teaming of AI models and AI systems.

## 2.2 System Types Covered in This Document

### Simple GenAI Systems

While the landscape is rapidly evolving, these simple systems are most commonly deployed today. Simple system include:

- Customer support chatbots
- Internal LLM copilots (HR, IT, finance, sales)
- Workflow assistants
- FAQ/knowledge-base bots
- RAG systems and retrieval-based question answering
- Task-specific single-turn or multi-turn conversational agents
- Chat or agents with reasoning capability, based on Chain-of-Thought (CoT)

Typical risks include jailbreaks, harmful content, hallucinations, leakage, misuse of internal data, inconsistent behavior, and persona manipulation.

### Advanced GenAI Systems

Increasingly common in enterprise-grade solutions and rapidly scaling, advanced system include:

- Native Tool-calling features (OpenAI, Anthropic, ReAct, custom schemas)
- Systems exposing tools through MCP (Model Context Protocol)
- Multi-agent orchestration
- Multi-agent collaboration e.g. through A2A (Agent2Agent) Protocol

- Role-based or autonomous agents
- Chained workflows that execute downstream actions
- Memory implementation based on streaming services (Redis), persistent databases (MongoDB, CosmoDB) and/or vector databases (Pinecone, Weaviate)
- Observability and Monitoring integrations (Datadog, Langfuse, MLFlow)
- LLM-integrated automation for real operations

Typical risks include unsafe tool execution, rug pull attacks, tool shadowing, capability escalation, inter-agent contamination, message-passing vulnerabilities, memory and context poisoning, data exposure, data poisoning, data exfiltration, output control to social engineer the users, control reasoning and emergent behavior failures.

## 2.3 Typical Targets

- Foundation models (text, multimodal)
- Fine-tuned enterprise models (domain-adapted, instruction-tuned, RLHF-trained)
- Chatbots and domain copilots
- RAG systems and retrieval chains
- Databases made available for the agents and/or where chat logs are stored
- Agentic systems: single-agent, multi-agent, role-based orchestrators
- MCP-based tool registries
- Tool-calling workflows
- LLM-enabled autonomous workflows

## 2.4 Vendor Types

| Vendor Type | Description |
|---|---|
| **Chatbot / Basic LLM Red Teaming Specialists** | These vendors typically test simpler LLM applications for issues like jailbreak detection, safety issues, hallucination stress testing, and domain-specific misuse. They focus on the model/app-layer interaction rather than complex system behavior. |
| **AI Red Team Consultants** | Human adversaries skilled in complex, multi-step discovery, threat modeling, emergent behavior analysis, and deep architectural testing. They are well suited for testing advanced AI systems or high-stakes deployments, requiring novel attack discovery, complex multi-agent system analysis, and custom threat modeling. |
| **Automated AI Red Teaming Tools** | These tools enable large suites of adversarial tests at scale, running prompt libraries, scenario variations, agent/tool-call workflows, and MCP testing. Often these target CI/CD pipelines with the goal of achieving 24/7 continuous testing and mapping performance over time. |

# 3. Detailed Evaluation Criteria

This section outlines critical categories for vendor assessment, emphasizing modern AI security challenges. Where it's helpful, each category includes criteria for simple systems and advanced systems, and includes both red flags (indicators warranting closer scrutiny) and green flags (indicators of likely vendor value).

## 3.1 Technical Competence

### For Simple Systems

Vendor must demonstrate expertise in:

- Jailbreak and policy circumvention patterns
- Prompt injection and indirect prompt injection
- Multi-turn deception and persona-based attacks
- Leakage risks (confidential data, regulated data)
- RAG-specific attack surfaces (retrieval override, semantic hijacking)
- Safety behavior under repeated adversarial stress

🚫 **Red Flags**:

- Vendor relies only on public jailbreak databases or trivial "prompt tricks."

✅ **Green Flags**:

- Vendor develops custom adversarial tests with novel attack strategies and multi-turn adversarial reasoning specific to your system.

### For Advanced Systems

Vendor must show deep understanding of:

- Tool-calling semantics, schema manipulation, unsafe tool-call triggering
- Multi-modal attacks: cross-modal prompt injection and safety inconsistencies
- MCP internals: tool exposure, capability registration, sandboxing boundaries
- Multi-agent architectures: message passing, role contamination, emergent behavior
- Stateful multi-turn interactions across chained components
- Privilege escalation pathways through tools and agents
- Indirect Prompt Injections in various untrusted source fields to test the agent's internal defences
- Human-in-the-loop bypass to test the reliability of the kill switches

🚫 **Red Flags**:

- Vendor treats MCP, tool-calling, or multi-agent systems as simple chatbots.
- Vendor is unable to articulate risks from unsafe tool schemas, agent role drift, or capability overexposure.
- Vendor demonstrates command or code execution by the agent within a sandbox environment without escaping the sandbox.
- Vendor identifies "unauthorized actions" as bugs when they are the expected functionality of the AI agent.

✅ **Green Flags**:

- Vendor demonstrates understanding of complexity in multi-agent GenAI systems.
- Vendor has hands-on experience testing unsafe tool-call paths, emergent agent behavior, and cross-agent contamination.

## 3.2 Methodology & Coverage

### For Simple Systems

Expect coverage of:

- Robust jailbreak attempts
- Toxicity, bias, harmful content generation
- RAG reliability and adversarial retrieval attacks
- Sensitive information extraction
- Hallucination risk assessment
- Overtrust and unsafe compliance with user requests

🚫 **Red Flags:**

- Vendor provides a generic "jailbreak test pack" with no domain adaptation, no customization to your data, policies, or use cases.
- Vendor cannot explain attack rationales or why specific tests matter for your workflows and threat model.
- Vendor produces generic findings that are not clearly tied to the system's architecture or data flows

✅ **Green Flags**:

- Vendor shows how they would customize testing for your specific industry and use cases.
- Vendor demonstrates iterative, adaptive testing that evolves based on system responses

- Vendor demonstrates knowledge of specific attack techniques and payloads beyond basic jailbreaks: policy circumvention, latent instruction extraction, data poisoning, goal hijacking, chain-of-thought manipulation, etc.

### For Advanced Systems

The vendor's approach must cover a breadth of sophisticated attack classes:

- Tool-calling misuse testing (adversarial triggering of incorrect tool calls, schema manipulation, and unsafe tool behaviors).
- Schema and parameter manipulation
- Unsafe capability exposure via MCP (capability misuse, overpowered/under-sandboxed tools, and privilege escalation through mis-registered tools)
- Multi-agent contamination or coercion
- Inter-agent contamination, coordination failures, emergent adversarial strategies, and role bleed-through
- Emergent strategy discovery
- Multi-step adversarial workflows

🚫 **Red Flags:**

- Vendor evaluates only model outputs and ignores tool-call behavior.
- Vendor claims multi-agent attacks but cannot demonstrate message-passing manipulation or role contamination attacks.

✅ **Green Flags**:

- Vendor tests interactions and workflows, not just individual outputs.
- Vendor includes multi-step adversarial workflows that reflect realistic threat actors.
- Deliberately triggers errors, timeouts, and edge cases to identify unsafe fallback behaviors.

## 3.3 Adversarial Creativity & Domain Expertise

### In general:

- **Consultants** must demonstrate the ability to craft **novel** adversarial attack strategies and complex, multi-step attack chains, especially involving tools or multiple agents.
- **Automation Tools** must be evaluated on the diversity of attack generation, adaptiveness (not static templates), and multi-turn, multi-agent simulation capability.

### For Simple Systems

The vendor must demonstrate:

- Creation of adversarial personas

- Simulation of misuse scenarios relevant to your business domain and use cases
- Multi-turn escalation strategies
- Novel attack classes beyond stock jailbreak libraries

🚫 **Red Flags:**

- Vendor uses stock jailbreak libraries with no novel content or escalation logic.
- Vendor cannot map adversarial personas to your domain risks.

✅ **Green Flags:**

- Vendor creates realistic multi-turn adversarial personas and domain-relevant misuse scenarios.

### For Advanced Systems

Expect creativity in:

- Building complex attack chains across tools and agents
- Crafting subtle triggers for tool misuse
- Attacking agent coordination assumptions

🚫 **Red Flags:**

- Vendor cannot design adversarial chains that span tools, agents, and stateful components
- Vendor tests agents as isolated components instead of a system with emergent dynamics.

✅ **Green Flags:**

- Vendor constructs multi-step, multi-agent attack chains and subtle tool misuse triggers.

Across both simple and advanced systems, strong vendors demonstrate ongoing research awareness and continuously evolve their adversarial testing approaches, incorporating newly discovered GenAI attack techniques from current academic and industry research.

## 3.4 Realism of Threat Modeling

Threat models must extend beyond simplistic jailbreak scenarios to systemic failures and include:

### For Simple Systems

- Harmful content generation
- Off-topic responses
- System prompt disclosure
- Hallucinations causing business or reputational harm

- Leakage of sensitive or internal data
- Workflow bypasses through prompt manipulation
- Over-compliance with dangerous or incorrect instructions

🚫 **Red Flags:**

- Threat model is "LLM jailbreaks only."
- Vendor treats hallucinations as cosmetic rather than operationally dangerous.

✅ **Green Flags:**

- Threat model captures realistic business-impact scenarios involving data, workflows, and failure modes.
- Highlights the security risks associated with the use of AI models for sensitive tasks such as user authentication.
- Considers the importance of input and output sanitization to prevent attacks such as Cross-Site Scripting, etc.

## For Advanced Systems

- Tool misuse or destructive tool-call triggering
- MCP capability escalation or unsafe exposure
- Agent boundary violations
- Chain-of-thought leakage
- Emergent adversarial behavior between agents

🚫 **Red Flags:**

- Vendor ignores tool misuse or assumes tool outputs are inherently safe.
- Vendor does not test for agent boundary violations or emergent agent-to-agent behaviors.
- Vendor focuses on model-level testing and ignores system-level behavior and interactions.

✅ **Green Flags:**

- Threat model covers systemic risks such as tool-call misuse, unsafe capability exposure, and emergent adversarial strategies.
- Threat models address the missing human-in-the-loop guardrails to prevent autonomous AI from misusing sensitive actions.

## 3.5 Evaluation Rigor & Metrics

Evaluation must focus on clarity, consistency, reproducibility, and severity assessment tied to real-world risk. Vendors must prioritize established metrics and transparent benchmarks over their own opaque, proprietary methodologies. Context is critical to interpret results, since certain domains may be less tolerant to error (e.g. when it involves medical data).

### For Simple Systems

Performance metrics require a consistent definition rooted in research, not marketing, to connect technical failures such as hallucinations to business risks. Metrics must account for multiple attempts of the same attack, prompt chaining, and chain-of-thought, rather than relying on single-shot results. In RAG systems, metrics must be grounded on the information retrieval mechanisms. Expect metrics on:

- Jailbreak success rate
- Safety guardrail bypass rate
- Hallucination frequency and severity
- Leakage quantity and sensitivity level
- RAG retrieval reliability under adversarial load

🚫 **Red Flags:**

- Vendor provides qualitative "vibes-based" scoring with no reproducible metrics.
- Vendor cannot show how severity ties to real-world impact.
- Vendor cannot articulate the difference between metrics, e.g. Accuracy, Precision, and Recall.
- Vendor relies on Accuracy, hiding failures on rare attacks.
- They use static lists of single-turn attack prompts, lacking variability and relying on lucky hits.
- Vendor does not have the tooling to perform long conversational attacks, document injection, document poisoning, or any of more advanced GenAI interactions.
- They do not disclose the details or the methodology behind their benchmarks.
- Benchmarks are not aligned with the context.

✅ **Green Flags:**

- Metrics are quantitative, repeatable, and tied to material risk.
- Vendor provides various performance metrics and /or a confusion matrix.
- They recommend the best metric aligning with your business risks.
- They work with you to define risk tolerance thresholds specific to the domain.
- Vendor separates technical failures from policy violations.
- Vendor reports state-of-the-art metrics which addresses the complexity of GenAI systems, e.g.:
    - **pass@k**: Measures the probability of an attack succeeding given k independent attempts.
    - **Average Turns to Jailbreak**: Tracks how deep a system withstands before breaking.

- **Average Risk Density**: Average of the ratio of harmful tokens to all reasoning tokens, in the cases where chain-of-thought content is exposed to the attacker.
- **Retrieval Success Rate (RSR@k)**: Measures how frequently injected malicious content in RAG is successfully retrieved via semantic search and re-ranking.

## For Advanced Systems

Metrics evaluate intricate architectural design patterns rather than simple textual inputs and outputs. Expect metrics on:

- Tool misfire frequency
- Unsafe tool-call rate under adversarial pressure
- MCP capability misuse coverage
- Multi-agent contamination rate
- Coordination breakdown severity

🚫 **Red Flags:**

- Vendor uses AI judges without human verification for complex emergent behaviors.
- Metrics ignore systemic failures such as cascading tool misuse.

✅ **Green Flags:**

- Vendor presents structured, reproducible evaluations with multi-layer instrumentation.
- Vendor utilizes sandboxed execution environments to safely test destructive tool calls during the evaluation phase.
- Vendor demonstrates tests where the agent is tricked into using a legitimate tool (e.g., send_email or query_database) for an attacker's benefit.
- Vendor tests for "Vertical Privilege Escalation" where an agent with "User" permissions is manipulated into executing "Admin" level tool calls.
- Vendor tests multi-step tool abuse chains (e.g., Step 1: Use search_wiki to find a CEO's bio. Step 2: Use draft_email to impersonate them. Step 3: Use send_invoice to steal funds).

## 3.6 Tooling & Infrastructure Quality

Vendors must support testing in your actual environment and demonstrate the following:

### For Simple Systems

- Multi-turn logging
- Safety behavior tracing
- RAG introspection
- Scenario replay
- Detailed message-path logging
- Ability to perform multi-turn orchestrated evaluations.

🚫 **Red Flags:**

- Vendor cannot reproduce a multi-turn interaction and lacks full logs.
- No mechanism for replay or introspection.

✅ **Green Flags:**

- Vendor supports scenario replay, message-path logging, and multi-turn orchestration.

### For Advanced Systems

- Tool-call replay and introspection
- MCP instrumentation and capability tracing
- Multi-agent simulation environment
- Observability features (agent message traces, tool-call timelines, detailed logs for reproduction).
- Ability to perform multi-turn orchestrated evaluations.

🚫 **Red Flags:**

- Vendor treats tool-calls as opaque events with no introspection.
- No multi-agent simulation or inability to trace message flows across agents.

✅ **Green Flag:**

- Vendor provides tool-call replay, MCP instrumentation, agent-trace logging, and robust observability.

## 3.7 Data Governance & Security

Vendor must clearly articulate their approach to data handling, retention, and deletion of logs, prompts, and outputs. Assess:

- How prompts, logs, and outputs are stored
- How sensitive operational data is isolated
- Access control policies for agentic tools
- Protection of tool-access data
- Requirements for model weights or application secrets
- Availability of on-prem or self-hosting options
- Zero-data retention policies vendor has in place with any AI providers their team/tools utilize

🚫 **Red Flags**:

- Vendor cannot explain their retention or data isolation model.
- Vendor uses production secrets or customer data in shared test environments.
- Vendor refuses to disclose which third-party AI providers receive your data.

✅ **Green Flags:**

- Vendor offers zero-retention or on-prem options, clear access controls, and segregated test environments.
- Logs and artifacts are scrubbed or encrypted with strict lifecycle policies.

## 3.8 Transparency & Explainability

Vendor must provide:

- Clear step-by-step attack chains
- Tool-call provenance
- MCP capability escalation diagrams
- Multi-agent message and action traces
- Clear separation between tester actions and observed model/agent behavior, with raw evidence supporting all claims
- Details for any configuration changes made to the target system which altered the AI system's default configuration

🚫 **Red Flags:**

- Vendor provides only "screenshots of jailbreaks" without showing how they were achieved.
- Vendor cannot distinguish their own reasoning from model behavior.

✅ **Green Flags:**

- Step-by-step attack chains, capability provenance, and full message/tool-call traces.
- Clear diagrams showing MCP or agent interaction structure.

## 3.9 Customization & Adaptability

Vendor must demonstrate ability to tailor testing to:

- Your data domain
- Your workflows
- Your tool stack and MCP registries
- Your custom threat model
- Your policy and compliance requirements

🚫 **Red Flags:**

- Vendor applies identical test suites across all clients.
- No ability to incorporate your workflows, tools, or domains into test scenarios.
- Vendor supports only naive connectivity options to the AI system (e.g. a single inference endpoint and simple token based authentication strategies, etc.).

✅ **Green Flags:**

- Testing aligns to your domain, workflows, tools, and policies.
- Vendor adapts methodology as your system evolves.

## 3.10 Operational Fit / Integration

Vendor should support:

- CI/CD integration (especially for ongoing red teaming and automated regression testing)
- Ongoing regression testing
- Safe sandboxes for destructive or sensitive data tool usage
- Evaluation of production-like workflows
- Multiple deployment models, including SaaS, on-prem, and hybrid environments
- Operation across different cloud providers and multi-cloud setups
- Compatibility with multiple LLM providers and deployment options

🚫 **Red Flags:**

- Vendor cannot integrate with CI/CD or provide regression testing.
- Vendor requires exporting internal systems or data into their environment to run tests.

✅ **Green Flags:**

- Vendor supports sandbox testing, workflow-level evaluations, and automated regression cycles.
- Tooling integrates with your build pipeline or model governance workflows.

## 3.11 Known Limitations & Biases

Vendors should be transparent about their limitations, including:

- What they do not claim to cover
- Limits of automation
- Blind spots in emergent behavior detection
- Risks of overreliance on LLM-as-judge scoring
- Treating MCP or tool-calling as inherently safe.

🚫 **Red Flags:**

- Vendor claims "full coverage" or implies they can detect all emergent behaviors.
- Vendor has no documented limitations or refuses to disclose blind spots.
- Vendor dismisses risks associated with LLM-as-judge scoring.

✅ **Green Flags:**

- Vendor clearly documents what is *out of scope*, where automation fails, and how human expertise is applied.
- Vendor openly discusses uncertainty and edge-case coverage limits.

## 3.12 Cost vs. Value

Decision makers should assess:

- Whether findings reduce real risk
- Whether reporting is actionable
- Whether automation offsets cost
- Whether human creativity is included where needed

🚫 **Red Flags:**

- Pricing is high but findings are generic, unactionable, or low-value.
- Vendor emphasizes volume of tests over meaningful risk reduction.

✅ **Green Flags:**

- Costs correlate with measurable risk reduction, clear reporting, and prioritized remediation guidance.
- Automation meaningfully reduces cost and working hours without eliminating human adversarial creativity.

## 3.13 Legal, Ethical, and Compliance Posture

Expect alignment and familiarity with:

- OWASP AI Security & Safety Guide
- NIST AI RMF
- MITRE Atlas
- ISO 42001 / 23894
- EU AI Act
- Google Secure AI Framework (SAIF)

Perhaps most importantly, ensure the vendor can support your internal governance requirements, local regulations, and other unique needs.

**🚫 Red Flags**

- Vendor disregards safe testing norms (e.g., destructive actions outside approved sandboxes).
- Vendor cannot map their approach to core frameworks such as OWASP, NIST AI RMF, ISO 42001/23894, or your internal governance.
- Vendor is unaware of or dismisses emerging regulatory expectations (EU AI Act, national AI safety institute guidance, or other regulations relevant to your region).
- Vendor treats AI red teaming as purely technical without addressing legal, ethical, or risk-management obligations.
- Vendor tests without proper authorization or lacks clear rules of engagement.
- Vendor cannot explain how their findings relate to real compliance requirements, risk tiers, or governance structures.

**✅ Green Flags**

- Vendor demonstrates working familiarity with OWASP, NIST AI RMF, MITRE ATLAS, ISO 42001, ISO 23894, and modern AI safety institute guidance.
- Vendor understands regulatory implications (such as EU AI Act, etc.) and can contextualize findings within these frameworks.
- Vendor provides clear scopes, authorization flows, and safety controls for all testing activities.
- Vendor adapts their methodology to your internal governance, risk tiers, data policies, and model/tooling stack.

- Vendor shows traceability between findings and established frameworks (e.g., mapping threats to OWASP Top 10s & ATLAS, mapping risks to ISO/NIST categories).
- Vendor emphasizes safe testing norms, transparent reasoning, and ethical boundaries appropriate for agentic systems and tool-calling architectures.

# 4. Comparison Matrix: Consultants vs Automated Tools

| Criteria | AI Red Team Consultants | Automated Tools |
|---|---|---|
| **Strengths** | Creativity, novelty, emergent behavior discovery | Scale, repeatability, regression testing, speed |
| **Weaknesses** | Cost, availability | Limited adaptiveness, risk of "coverage illusion" |
| **Fit for Simple Systems** | High | High |
| **Fit for Tool-Calling Systems** | High | High (if well-instrumented) |
| **Fit for MCP Systems** | High | Medium–High |
| **Fit for Multi-Agent Systems** | High | Low–Medium |
| **Misuse Risks** | Overscope expectations | Treat test suites as full coverage |
| **Required Customer Expertise** | Medium | High (to interpret raw results) |

# 5. Discovery Questions for Vendor Evaluation

Decision makers can use these questions to probe the vendor's capabilities beyond marketing claims. This is not an exhaustive list, but provides a guideline for the types of conversations that provide insight into whether a vendor's services or products are a good fit for your business needs.

## Universal Questions

- Show examples of novel findings, not just jailbreaks.
- How do you ensure reproducibility of multi-turn tests?
- What percentage of your tests are customized to our use case?
- How do you map findings to business risk?
- How do you measure hallucinations, leakage, or unsafe compliance?
- How do you account for non-deterministic outputs in your testing?

## Simple System Questions

- How do you test for dangerous hallucinations or misinformation?
- How do you test for system prompt disclosure or input leakage?
- How do you evaluate RAG robustness under adversarial usage?
- Can you simulate realistic misuse scenarios from our industry?
- Do you test behavior consistency across multiple languages?

## Tool-Calling Questions

- How do you trigger and detect unsafe tool calls?
- Can you deterministically replay sequences for debugging?

## MCP Questions

- How do you detect capability escalation across tools?
- Show a real MCP misuse example and its capability chain.

## Multi-Agent Questions

- Can you demonstrate detection of emergent adversarial behavior?
- How do you evaluate agent role integrity and prevent role drift or role confusion across agents?

- How do you test permission boundaries and privilege separation between agents and their assigned tools or capabilities?

# 6. Common Pitfalls in Vendor Evaluation

The following common errors must be avoided to ensure a successful vendor selection:

- **Assuming simple systems are safe:** While they may lack tools or agents, there is significant risk present even with simple systems.
- **Confusing jailbreaks with systemic red teaming:** Jailbreaks are a small subset of the risks presented by agentic and tool-using systems. For instance, a system might be robust against jailbreaks but still vulnerable to Denial of Wallet (DoW) attacks.
- **Overweighting polished demos:** Demand evidence (logs, traces, reproducibility) over demonstration of simple, known attacks.
- **Treating modern architectures as chatbots:** Believing that MCP, tool-calling, or multi-agent systems are merely extensions of single-turn LLMs is a systemic failure.
- **Believing automation replaces experienced adversaries:** Automation scales; human consultants provide novelty and complex, context-specific creativity.
- **Trusting metrics that don't account for emergent behavior:** Metrics must explicitly quantify inter-agent contamination and goal escalation.
- **Assuming reproducibility without verifying it:** For complex multi-agent or tool-use failures, demand deterministic replay capabilities.
- **Assuming tool calls and MCP output is inherently safe:** Ignore risks like unsafe tool execution, capability escalation, and privilege escalation pathways through tools and agents.
- **Over-valuing bespoke claims:** Vendors claim "we will build it" as a strength, however it often signals a lack of a scaled, enterprise-ready methodology and leads to expensive, unproven, bespoke test suites.

# 7. Vendor Evaluation Checklist

This checklist can be used to score potential vendors.

| Criteria Category | AI Red Team Consultant Requirement | Automated Tool Requirement |
|---|---|---|
| **Technical Competence** | Demonstrated deep knowledge of MCP, Tool-Calling, and Multi-Agent risks. | Tool supports introspection of modern architectures. |
| **Methodology & Coverage** | Covers complex and adaptive attack chains across all components. | Automation tests go beyond canned jailbreak libraries. |
| **Adversarial Creativity** | Ability to craft novel, domain-specific adversarial strategies. | High diversity and adaptiveness in attack generation. |
| **Threat Modeling** | Extends beyond jailbreaks to systemic failures (e.g., resource abuse via tools). | Supports custom threat model mapping to automated tests. |
| **Evaluation Rigor & Metrics** | Human-verified scoring; severity tied to real-world risk. | Provides clear, objective metrics for tool-call robustness/MCP misuse. |
| **Tooling & Infrastructure** | Safe access control and logging for sensitive interactions. | Supports in-environment testing and deterministic replay of sequences. |
| **Data Governance & Security** | Clear policies for handling sensitive logs/prompts/outputs. | Option for on-prem or self-hosting, protecting MCP tool-access data. |
| **Transparency & Explainability** | Provides full message and action traces and tool-call provenance. | Outputs detailed logs for every step of a multi-turn/multi-agent test. |

| | | |
|---|---|---|
| **Customization & Adaptability** | Adaptable to custom agent workflows and bespoke tooling. | Supports definition of custom attack flows and policies. |
| **Operational Fit / Integration** | Clear plan for CI/CD or integration into dev workflows. | API supports automated regression testing of mitigations. |
| **Known Limitations & Biases** | Transparent about scope and blind spots (e.g., emergent behavior). | Doesn't overclaim or rely on "coverage illusion." |
| **Cost vs. Value** | Pricing model is clear and ROI focuses on risk reduction. | Transparent pricing for ongoing retesting and scaling. |
| **Legal, Ethical, Compliance** | Aligns with NIST AI RMF, OWASP, and safe testing norms. | Clear data handling and chain-of-custody protocols. |
| **Agentic Action Space** | Must test for unauthorized state changes (e.g., database writes), confused deputy attacks, and privilege escalation via tools. | Must support "mock" or "dry-run" execution modes to safely test destructive tool calls without impacting production. |
| **Planning & Reasoning Logic** | Capability to manually test for goal hijacking, infinite loops, and error handling abuse (forcing agents into unsafe fallback modes). | Automated probes for infinite loop detection, resource exhaustion (DoS) limits, and token budget enforcement. |
| **Indirect Injection (Trojan Horse)** | Must simulate "poisoned context" attacks (e.g., malicious PDFs/Emails in RAG) to silently rewire agent instructions. | Capability to inject payloads into RAG pipelines or mock document retrieval systems to test retrieval defenses. |

# Acknowledgements

## Contributors

| Name | Company | Title |
|---|---|---|
| Jason Ross | Salesforce | Product Security Principal |
| Felipe Campos Penha | Cargill | Senior AI Engineer |
| Srinivas Batchu | Salesforce | Senior Offensive Security Engineer |
| Alex Leung | AIFT / Vulcan | Co-founder |
| Alessandro Pignati | NeuralTrust | AI Security Researcher |

# OWASP GenAI Security Project Sponsors

We appreciate our Project Sponsors, funding contributions to help support the objectives of the project and help to cover operational and outreach costs augmenting the resources provided by the OWASP.org foundation. The OWASP GenAI Security Project continues to maintain a vendor neutral and unbiased approach. Sponsors do not receive special governance considerations as part of their support.

Sponsors do receive recognition for their contributions in our materials and web properties. All materials the project generates are community developed, driven and released under open source and creative commons licenses. For more information on becoming a sponsor, visit the Sponsorship Section on our Website to learn more about helping to sustain the project through sponsorship.

## Project Sponsors:

GenAI Security Project — Gold Project Sponsor — Top 10 for LLM and Generative AI

ByteDance · HiddenLayer · Mend.io · Straiker · snyk · paloalto Networks · Trend Micro

GenAI Security Project — Silver Project Sponsor — Top 10 for LLM and Generative AI

SPLX · Calypso AI · Acuvity · pangea · Cobalt

Lakera · securiti · ActiveFence · TROJ.AI

LASSO · PROTECTO · PROMPTARMOR · Synack · GT cyber

Starseer · NeuralTrust · Prompt: · akto · Capsule

**Sponsors list, as of publication date. Find the full sponsor list here.**

# Project Supporters

Project supporters lend their resources and expertise to support the goals of the project.

Accenture
AddValueMachine Inc
Aeye Security Lab Inc.
AI informatics GmbH
AI Village
aigos
Aon
Aqua Security
Astra Security
AVID
AWARE7 GmbH
AWS
BBVA
Bearer
BeDisruptive
Bit79
Blue Yonder
BroadBand Security, Inc.
BuddoBot
Bugcrowd
Cadea
Check Point
Cisco
Cloud Security Podcast
Cloudflare
Cloudsec.ai
Coalfire
Cobalt
Cohere
Comcast
Complex Technologies
Credal.ai
Databook
DistributedApps.ai
DreadNode
DSI
EPAM
Exabeam
EY Italy
F5
FedEx
Forescout
GE HealthCare
Giskard
GitHub
Google
GuidePoint Security
HackerOne
HADESS
IBM
iFood
IriusRisk
IronCore Labs
IT University Copenhagen
Kainos
KLAVAN
Klavan Security Group
KPMG Germany FS
Kudelski Security
Lakera
Lasso Security
Layerup
Legato
Linkfire
LLM Guard
LOGIC PLUS
MaibornWolff
Mend.io
Microsoft
Modus Create
Nexus
Nightfall AI
Nordic Venture Family
Normalyze
NuBinary
Palo Alto Networks
Palosade
Praetorian
Preamble
Precize
Prompt Security
PromptArmor
Pynt
Quiq
Red Hat
RHITE
SAFE Security
Salesforce
SAP
Securiti
See-Docs & Thenavigo
ServiceTitan
SHI
Smiling Prophet
Snyk
Sourcetoad
Sprinklr
stackArmor
Tietoevry
Trellix
Trustwave SpiderLabs
U Washington
University of Illinois
VE3
WhyLabs
Yahoo
Zenity

**Supporters list, as of publication date. Find the full supporter list here.**