

LLM Security Testing

Generative AI is rapidly transforming how organizations build products, automate workflows, support employees, and interact with customers. But as large language models become integrated into applications, data pipelines, APIs, plugins, and autonomous agent workflows, they introduce a new and complex attack surface that traditional security testing may not fully address.

LLM-enabled systems can behave unpredictably when exposed to adversarial input, untrusted data, or manipulated context. Attackers may attempt to bypass safety controls, extract sensitive information, manipulate outputs, abuse connected tools, or influence the system into taking unintended actions. Risks such as prompt injection, jailbreaks, excessive agency, insecure output handling, sensitive information disclosure, and retrieval-augmented generation weaknesses can all create meaningful business, operational, and data protection concerns.

Bitdefender Offensive Services provides customized LLM Security Assessments to help organizations identify, validate, and remediate AI-specific vulnerabilities before they can be exploited. Each engagement is tailored to the design, logic, and business purpose of your AI implementation, giving you a clear understanding of where your LLM applications are exposed and how to strengthen them over time.

What We Assess

Every AI implementation is different. Bitdefender tailors each assessment to your specific architecture, use case, integrations, and risk profile of your AI implementation.

Testing may include evaluation of:

- ↳ Prompt injection, jailbreak resilience, and safety control bypass
- ↳ System prompt leakage and sensitive information disclosure
- ↳ Excessive agency, unintended tool use, and connected system abuse
- ↳ Retrieval-augmented generation, vector store, and context manipulation weaknesses
- ↳ Plugin, API, connector, and data flow trust boundary risks
- ↳ Logging, monitoring, detection, response, and AI business logic gaps

Our approach assesses the full LLM application ecosystem, including application logic, data sources, orchestration layers, connected services, and user interaction points.

At-a-Glance

Bitdefender's LLM Security Assessment evaluates AI-enabled systems under realistic adversarial conditions.

Using automated testing, manual adversarial techniques, and controlled exploitation, our consultants assess how your LLM applications process input, generate responses, and interact with internal data, external tools, APIs, and business workflows to identify gaps that may allow attackers to manipulate model behaviour or abuse connected systems.

The assessment delivers practical, evidence-based insight into your AI security posture — helping security, engineering, and risk teams understand vulnerability impact and prioritize remediation.

Key Benefits

- ↳ **Gain visibility into AI-specific risk** across prompts, interfaces, data pipelines, APIs, plugins, agents, and connected workflows.
- ↳ **Identify vulnerabilities before attackers do** like prompt injection, jailbreak, data leakage, excessive agency, insecure output, and tool abuse risks.
- ↳ **Validate AI deployment security** for production, expansion, or wider business adoption through automated scanning and expert-led manual adversarial testing.
- ↳ **Safely understand real-world business impact** within agreed scope to demonstrate practical risk to data, systems, users, and operations.
- ↳ **Strengthen secure design decisions** - by understanding how models interact with applications, users, tools, and data sources.
- ↳ **Governance, audit, and compliance support** with severity ratings, evidence, reproduction steps, and remediation guidance.

Our Methodology

Bitdefender combines automated coverage with expert-led manual adversarial testing to identify weaknesses that tools alone may miss. Each assessment is tailored to your AI architecture, use case, integrations, and risk profile.

Our testing process follows four key phases:

Reconnaissance and Scoping

We review your AI architecture, data flows, model interfaces, prompts, trust boundaries, connected APIs, plugins, agents, and retrieval systems to define a focused test scope.

Adversarial Testing

We simulate realistic attacks to assess how the LLM application responds to manipulation, untrusted input, prompt injection, jailbreak attempts, data extraction, insecure outputs, and unintended tool use.

Controlled Exploitation

Where vulnerabilities are identified, we safely validate their potential impact within the agreed scope, helping determine whether weaknesses could lead to data exposure, unauthorized actions, or business logic abuse.

Reporting and Remediation

We provide risk-rated findings, supporting evidence, reproduction steps, business impact analysis, and prioritized remediation guidance mapped to recognized AI security frameworks where appropriate.

Secure Your AI Systems Before Attackers Do

By combining offensive security expertise with a practical understanding of how generative AI applications are built, integrated, and attacked, Bitdefender helps organizations identify relevant, reproducible, and actionable vulnerabilities. The LLM Security Assessment gives security, engineering, governance, and risk teams the insight needed to validate controls, strengthen AI deployments, reduce the risk of misuse or data exposure, and adopt generative AI with greater confidence.

Contact Us

For more information on Bitdefender Phishing, [please reach out to us](#).

Bitdefender is a cybersecurity leader delivering best-in-class threat prevention, detection, and response solutions worldwide. Guardian over millions of consumer, business, and government environments, Bitdefender is one of the industry's most trusted experts for eliminating threats, protecting privacy and data, and enabling cyber resilience. With deep investments in research and development, Bitdefender Labs discovers over 400 new threats each minute and validates around 40 billion daily threat queries. The company has pioneered breakthrough innovations in antimalware, IoT security, behavioral analytics, and artificial intelligence, and its technology is licensed by more than 150 of the world's most recognized technology brands. Launched in 2001, Bitdefender has customers in 170+ countries with offices around the world.

Release Date: June 2026

For more information, visit <https://www.bitdefender.com>.

All Rights Reserved. © 2026 Bitdefender.

All trademarks, trade names, and products referenced herein are the property of their respective owners.

Romania HQ

Orhideea Towers
15A Orhideelor Road,
6th District,
Bucharest 060071
T: +40 21 4412452

6301 NW 5th Way,
#4300,
Fort Lauderdale,
FL, 33309
T: +1 954-776-6262