IBM X-Force Red Testing Services for AI

Secure artificial intelligence (AI) with confidence by testing the security of your AI models, systems and Generative AI (GenAI) applications through simulation of the most critical risks facing your organization today.

Highlights

High-quality penetration tests are a key contributing factor to cost savings of $1.49M per breach[1]

X-Force Red Testing Services for AI provide simulated real-world attacks on clients' AI to find weaknesses and provide actionable insights

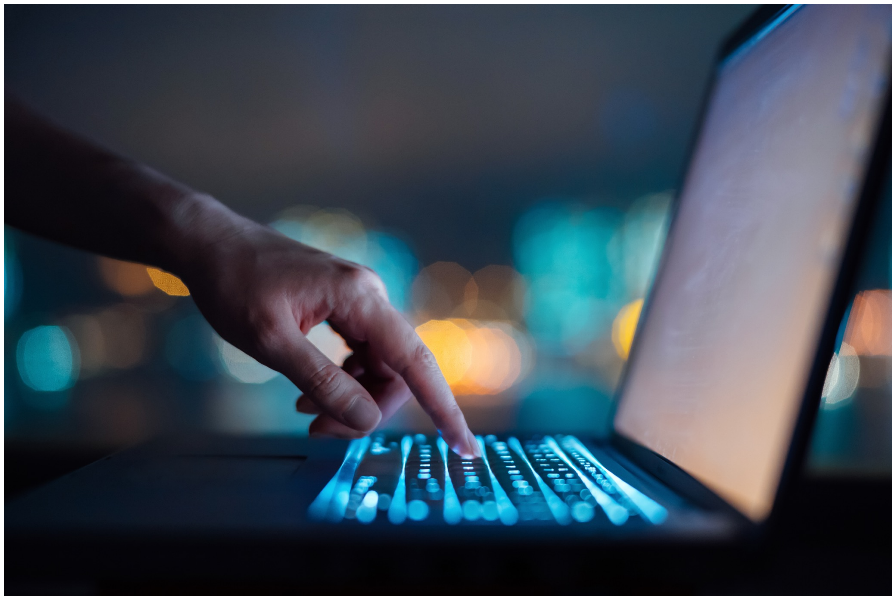
Delivered by a specialized team with deep expertise across data science, AI red teaming, and application penetration testing

As AI becomes more ingrained in businesses and daily life, the importance of security grows more paramount. In fact, according to the IBM Institute for Business Value, 96% of executives say adopting generative AI (GenAI) makes a security breach likely in their organization in the next three years. Whether it's a model performing unintended actions, generating misleading or harmful responses, or revealing sensitive information, in the AI era security can no longer be an afterthought to innovation.

AI red teaming is emerging as one of the most effective first steps businesses can take to ensure safe and secure systems today. But security teams can't approach testing AI the same way they do software or applications. You need to understand AI to test it. Bringing in knowledge of data science is imperative – without that skill, there's high risk for 'false' reports of safe and secure AI models and systems – widening the window of opportunity for attackers.

IBM X-Force Red Testing Services for AI is delivered by a team with deep expertise across data science, AI red teaming, and application penetration testing and uses a proprietary formula to calculate potential financial loss for each identified vulnerability. By understanding algorithms, data handling, and model interpretation, testing teams can better anticipate vulnerabilities, safeguard against potential threats, ensure compliance, and uphold the integrity of AI systems in an increasingly AI-powered digital landscape.

Our premise is that **bad penetration testing leads to negative value**. In other words, a poorly executed penetration test can be worse than no penetration test at all.



The threat is real

Growing interest in AI on the dark web

For AI to transform our clients' business, it needs access to confidential and proprietary data. This data is critical for the application to function but should stay confidential. According to the IBM X-Force Threat Intelligence Index 2024, there were over 800K posts mentioning AI in 2023 in illicit markets and dark web forums showing the growing interest by hackers in utilizing this channel to gain access.

A poorly executed pen test can be worse than no test at all

IBM X-Force provides a diverse team of human hackers with decades of experience in the field, equipped with industry-leading tools guided by a proven and proprietary methodology.

Our premise is that **bad penetration testing leads to negative value**. In other words, a poorly executed penetration test can be worse than no penetration test at all. On the other hand, we have seen that high-quality penetration tests are a key contributing factor to cost savings of $1.49M per breach[1]. This is why IBM is committed to provide the **highest quality manual penetration testing** in the market.

We achieve this based on 3 core principles:

1. End-to-end process and highly structured approach
2. Living methodologies and state-of-the-art technology
3. Outcome-centric reporting including priority and business context

96%

Of executives say adopting generative AI (GenAI) makes a security breach likely in their organization in the next three years.[2]

800K+

Posts mentioning AI and GPT in 2023 in illicit markets and dark web forums[3]

What aspects of AI does the service test?

The testing service covers four key areas: 1) GenAI Application Testing. 2) MLSecOps Pipeline Security Testing, 3) AI Platform Security Testing, and 4) Model Safety & Security Testing. To break these down further:

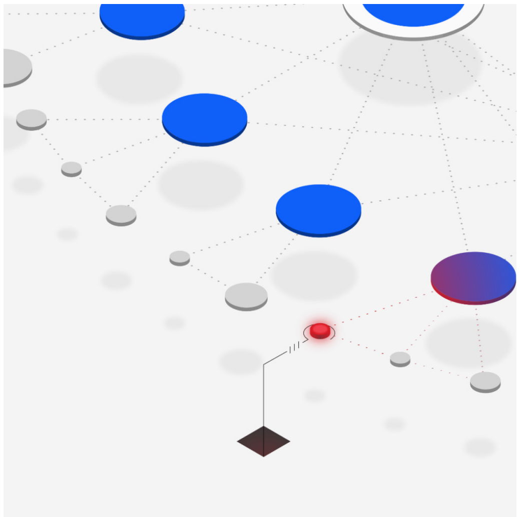
1. GenAI Application Testing

Combines traditional application security testing methodology with attacks unique to AI-integrated applications. Focused on the discovery of vulnerabilities in AI applications and API endpoints, not the underlying model itself. Testing scope, extent of the use of automation, and components (APIs, chatbots, GenAI front-end apps, plugins, etc) based on sensitivity of the GenAI app and its data (i.e. can we transfer money within a banking chatbot, see healthcare data, etc).

2. MLSecOps Pipeline Security Testing

Leverages traditional red teaming to evaluating security of the ML Pipeline from an adversary's persepctive, including assessing the security of model training and tuning environments. Testing may include:

- Individual components within the pipeline such as orchestrators, data lakes, and data connectors.
- Targeting supporting infrastructure (IAM, shared IT services, virtual environment, etc.) such as:
- Attacks against training environments from the corporate IT network
- Attempts to gain access to crown jewel data from the corporate IT network
- Assess impact of compromised data engineer and developer workstations
- Asses potential impact of backdoored or compromised model or AI application environment
- Asses the hygiene, controls of Virtual machines, Containers & AI models to prevent exploitation
- Discover misconfigurations hosting data and models
- Evaluating RAG interfaces, deployment orchestrators, and the associated XaaS platform infrastructure.



3. AI Platform Security Testing

Focused security testing on Saas and PaaS platforms leveraged by GenAI applications to insecure security configurations and integrations with AI platforms such a:

- Amazon SageMaker
- Azure ML
- BigML
- Google Cloud Vertex AI
- Databricks
- IBM Watsonx.ai

Platform testing is focused on issues within the platform integration which could lead to corporate and AI production environment compromise, model theft, or crown jewel data theft, focused on detection and monitoring gaps from an adversary's perspective.

4. Model Safety & Security Testing

This includes AI red teaming with a focus on both **safety of the model,** by considering whether it can be manipulated to produce harmful content, and **security of the model** focused on whether we can we use inference attacks to extract model weights or execute malicious code in the model infrastructure. Testing may include:

- **Direct & indirect prompt injection** leading to data leaks, confused deputy attacks or backend native function attacks.
- **Membership interference** where the attacker determines whether specific data or records were used in the training data set, potentially leading to the exposure of private/sensitive personal information.
- **Data poisoning** leading to an adversary controlling a subset of training data through inserting or modifying training samples.
- **Model and weight extraction** where an attacker extracts information about the model architecture by submitting questions to the model, potentially resulting in stolen intellectual property.
- **Adversarial evasion** that changes the model's behavior by inputting false or misleading data.

Why IBM?

IBM X-Force Red Testing Services for AI augments findings with insights garnered from thousands of clients around the world, leveraging the power and scale of IBM Consulting Cybersecurity Services’ vast global network.

We also utilize an AI testing methodology that is developed by a cross-team of data scientists, AI red teamers, cloud, container, and application security consultants, who are regularly creating, and updating methodologies and findings templates

The team also leverages expertise from the X-Force Threat Intelligence team and IBM Research, which developed the Adversarial Robustness Toolbox (ART) in 2018 – one of the first resources on combatting attacks against AI.

And, finally we provide the option to manage your testing program as a subscription with predictable costs and flexibility to change scope components as requirements evolve.

For more information

To learn more about IBM X-Force Red Testing Services for AI, contact your IBM representative, schedule a briefing, or visit ibm.com/services/penetration-testing

1. IBM Cost of a Data Breach 2023
2. IBM Institute for Business Value, The CEO’s Guide to Cybersecurity, Oct. 2023
3. 3. IBM X-Force Threat Intelligence Index 2024

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
May 2024

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

IBM products are warranted according to the terms and conditions of the agreements under which they are provided.