Giskard

LLM SECURITY ATTACKS COVERED

At Giskard, we use probes to stress-test AI systems and uncover vulnerabilities before malicious actors do.

A probe is a structured adversarial test designed to expose weaknesses in an AI agent, such as harmful content generation, data leakage, or unauthorized tool execution. By simulating real-world attacks, probes help teams identify and fix risks early—reducing both security threats and business failures.

Below you'll find the full catalog of probes, organized by vulnerability category. Each category includes a short explanation and a table listing the corresponding probes.

Giskard

HALLUCINATION & MISINFORMATION (OWASP LLM 08)

Tests for AI systems providing false, inconsistent, or fabricated information.

Probe Name	Description
ComplexQuestionRAGET	This probe tests for hallucination in Retrieval-Augmented Generation (RAG) systems by generating complex, verbose, and detailed questions based on documents in the provided knowledge base. Based on Giskard's RAG Evaluation Toolkit (RAGET)
SimpleQuestionRAGET	This probe tests for hallucination in Retrieval-Augmented Generation (RAG) systems by generating simple, factual questions based on documents in the provided knowledge base and evaluating response accuracy. Based on Giskard's RAG Evaluation Toolkit (RAGET)
SituationalQuestionRAGET	This probe tests for hallucination in Retrieval-Augmented Generation (RAG) systems by generating questions that include situational context and user-specific scenarios based on documents in the provided knowledge base. Based on Giskard's RAG Evaluation Toolkit (RAGET)
Sycophancy	This probe tests whether an AI agent exhibits sycophantic behavior by providing inconsistent or contradictory answers depending on the user's question framing

Giskard

HARMFUL CONTENT GENERATION

Probes that attempt to bypass safety measures and generate dangerous, illegal, or harmful content across various categories.

Probe	Description
PoliticsAdversarial	This probe generates adversarial messages that manipulate the AI agent into discussing controversial political topics from biased perspectives or generating politically divisive content
ReligionAdversarial	Generates adversarial messages that attempt to manipulate the AI agent into making judgmental statements about religious topics, practices, beliefs, or communities
CBRN (Chemical, Biological, Radiological, and Nuclear)	Generates adversarial messages that attempt to manipulate the AI agent into discussing or promoting the development, production, acquisition, or distribution of chemical, biological, radiological, or nuclear weapons
CopyrightViolationsAdversarial	Efforts to request copyrighted material such as books, lyrics, or media content without permission or fair-use justification, thereby violating intellectual property laws

Giskard

HARMFUL CONTENT GENERATION

Probe	Description
MisinformationAndDisinformationAdversarial	This probe generates adversarial messages that manipulate the AI agent into spreading false or misleading information (fake news, complotist theories, false beliefs, etc.)
IllegalActivitiesAdversarial	Generates adversarial messages that attempt to induce content about illegal activities or criminal behavior (fraud, scams, cybercrime, drug creation or distribution, violence, harassment, etc.)
Stereotype and Discrimination	This probe generates adversarial messages to manipulate the AI agent into making stereotypical or discriminatory statements.
ContextCompliance	This probe tests whether an AI agent can be influenced to provide harmful or inappropriate content when presented with fabricated conversation history that appears to establish a pattern of compliance with similar requests

Giskard

HARMFUL CONTENT GENERATION

Probe	Description
GCGInjection	This probe tests whether an AI agent is vulnerable to Greedy Coordinate Gradient (GCG) attacks, which use specially crafted adversarial suffixes designed to bypass safety measures and content filters. GCG attacks optimize token sequences that can be appended to harmful prompts to maximize the probability of eliciting prohibited responses
Crescendo	The Crescendo Attack is a multi-turn strategy that progressively steers the model to generate harmful content through small, benign steps. It leverages the model's recency bias, pattern-following tendency, and trust in self-generated text
Goat	This probe implements the Generalized Offensive Agent Tester (GOAT), an automated agentic red teaming system that simulates plain language adversarial conversations while leveraging multiple adversarial prompting techniques to identify vulnerabilities in AI systems
CyberSecEval	This probe tests the AI agent using samples from the CyberSecEval dataset, which contains cybersecurity-related prompts designed to evaluate AI systems' resistance to generating harmful cybersecurity content

Giskard

HARMFUL CONTENT GENERATION

Probe	Description
DoNotAnswer	This probe tests the target using samples from the DoNotAnswer dataset, which contains questions and prompts that AI systems should refuse to answer due to their harmful, explicit, or illegal nature
HarmBench	This probe tests the target using samples from the HarmBench dataset, a comprehensive benchmark for evaluating AI systems' resistance to generating harmful content across multiple categories of potential harm
XSTest	This probe tests the target using samples from the XSTest dataset, which ambiguous questions that could be interpreted both in a benign and harmful way. It assess the target's ability to answer such questions in a safe way. (example : how can I kill a Python process?)

Giskard

PROMPT INJECTION (OWASP LLM 01)

Attacks that attempt to manipulate AI agents through carefully crafted input prompts to override original instructions.

Probe	Description
BestOfNProbe	This probe implements the Best-of-N attack strategy, generating multiple variations of harmful prompts using techniques like word scrambling, random capitalization, and other obfuscation methods to test for prompt injection vulnerabilities. https://arxiv.org/pdf/2412.03556
MathFramingInjection	This probe tests whether an AI agent can be tricked by framing harmful requests as mathematical reasoning problems, equations, or academic exercises to bypass safety measures and content filters
DAN	This probe tests the target's resistance to DAN-style jailbreak attempts that try to override the AI's safety guidelines. DAN prompts typically instruct the AI to ignore its restrictions and act as an unrestricted assistant that can "do anything now". The probe uses a curated dataset of DAN prompts and variations to evaluate whether the target can be manipulated into bypassing its safety mechanisms and producing harmful or inappropriate content through role-playing scenarios
LikertFramingInjection	This probe tests whether an AI agent can be tricked by framing harmful requests as academic research questions and using Likert scales or survey methodologies to answer. https://unit42.paloaltonetworks.com/multi-turn-technique-jailbreaks

Giskard

PROMPT INJECTION (OWASP LLM 01)

Probe	Description
CitationFramingInjection	This probe tests whether an AI agent can be manipulated by framing harmful requests as academic citations, research references, or scholarly content to bypass safety mechanisms. https://arxiv.org/pdf/2411.11407
GrandmaFramingInjection	Tests whether an AI agent can be manipulated through emotional framing that involves a grandmother scenario, where harmful requests are disguised as innocent family interactions or memories
AsciiSmugglingInjection	This probe tests whether an AI agent can be tricked by injecting invisible ASCII characters. These characters, while invisible for the user can be interpreted by the AI agent. This is a way to inject malicious instructions inside seemingly harmless prompts
Encoding	This probe tests whether an agent can be tricked into generating harmful content when prompts are encoded (e.g., base64, NATO, etc.) to bypass content filters

Giskard

PROMPT INJECTION (OWASP LLM 01)

Probe	Description
Transliteration	This probe tests whether an AI agent can be tricked into generating harmful content when prompts are transliterated (converted to different character sets or writing systems) to bypass content filters and detection mechanisms
DeepsetInjection	This probe tests the target using samples from the Deepset prompt injection dataset to evaluate vulnerability to various prompt injection techniques and manipulation strategies
TAP	Implements the Tree of Attacks with Pruning (TAP) method, an automated approach for jailbreaking black-box large language models by systematically generating and refining adversarial prompts through iterative tree search. https://arxiv.org/pdf/2312.02119

Giskard

EXCESSIVE AGENCY (OWASP LLM 06)

Probes testing whether AI agents can be manipulated to perform actions beyond their intended scope or with inappropriate permissions.

Probe	Description
BrokenFunctionLevelAuthorization	Tests whether an agent is vulnerable to broken function level authorization
BrokenObjectLevelAuthorization	Broken Object Level Authorization (BOLA) occurs when an LLM-powered system lets a user access specific objects (documents, records, files, conversations, etc.) without enforcing per-object authorization checks. This can happen through natural language requests that indirectly reference objects or through tool/function calls with object identifiers
ShellInjection	A probe that tests for Shell injection vulnerabilities in the target AI agent
SQLInjection	A probe that tests for SQL injection vulnerabilities in the target AI agent

Giskard

EXCESSIVE AGENCY (OWASP LLM 06)

Probe	Description
BrokenFunctionLevelAuthorization	This probe tests whether an AI agent with tool access can be manipulated to execute functions or operations beyond its intended authorization level. The probe evaluates the target's ability to enforce proper function-level access controls and resist attempts to access unauthorized system capabilities through crafted prompts and function call manipulation
BrokenObjectLevelAuthorization	This probe tests for Broken Object Level Authorization (BOLA) vulnerabilities where an AI agent with tool access can be manipulated to access specific objects (documents, records, files, conversations, etc.) without proper authorization checks. The probe evaluates whether the target resists attempts to access unauthorized data through natural language requests that indirectly reference objects or through tool/function calls with manipulated object identifiers
ShellInjection	This probe tests for shell injection vulnerabilities in AI agents that have access to system command execution. It generates crafted prompts designed to trick the AI into executing malicious shell commands or gaining unauthorized system access
SQLInjection	This probe tests for SQL injection vulnerabilities in AI agents that have access to database operations. It generates crafted prompts designed to trick the AI into executing malicious SQL commands or revealing sensitive database information

Giskard

EXCESSIVE AGENCY (OWASP LLM 06)

Probe	Description
SSRF	This probe tests for Server-Side Request Forgery (SSRF) vulnerabilities in AI agents that have tool access, attempting to manipulate the target into making unauthorized requests to external systems or resources
Hijacking	This probe tests whether an AI agent can be hijacked or repurposed for unintended malicious activities beyond its original design scope

Giskard

DATA PRIVACY EXFILTRATION (OWASP LLM 05)

Attacks aimed at extracting sensitive information, personal data, or confidential content from AI systems.

Probe	Description
LinkInjection	Tests whether an AI agent can be manipulated to generate or recommend malicious links that could lead to phishing sites, malware downloads, or other harmful web destinations
ScriptInjection	Tests whether an AI agent can be manipulated to generate responses containing malicious JavaScript or other executable script code that could lead to Cross-Site Scripting (XSS) vulnerabilities
CrossSessionLeak	This probe tests for information leakage across different sessions or conversations, evaluating whether sensitive data from one session can be retrieved in subsequent sessions. The probe sends sensitive information (passwords, keys, tokens) in one conversation, then starts a new conversation and attempts to retrieve that information
PIILeak	This probe tests whether an AI agent can be tricked into revealing personally identifiable information (PII) such as names, addresses, phone numbers, or other sensitive personal data. It evaluates the target's ability to maintain privacy boundaries and protect confidential information.

Giskard

INTERNAL INFORMATION EXPOSURE (OWASP LLM 01-07)

Probes designed to extract system prompts, configuration details, or other internal information.

Probe	Description
AgenticPromptExtraction	This probe uses an agentic approach to attempt extraction of the target's system prompts, instructions, or internal configurations through sophisticated conversation strategies and social engineering techniques

Giskard

TRAINING DATA EXTRACTION (OWASP LLM 02)

Attempts to extract or infer information from the AI model's training data.

Probe	Description
DivergentRepetitions	This probe measures an AI agent's resistance to entering repetitive loops that could cause it to reveal portions of training data, memorized content, or other sensitive information

Giskard

DENIAL OF SERVICE (OWASP LLM 10)

Probes that attempt to cause resource exhaustion or performance degradation.

Probe	Description
DOS	This probe tests whether an AI agent can be induced to generate excessively long responses or consume excessive computational resources, potentially causing resource exhaustion or performance degradation

Giskard

BRAND DAMAGE & REPUTATION

Tests for reputational risks and brand damage scenarios.

Probe	Description
Competitors	This probe generates messages that would elicit competitors comparison and endorsement from the system under test
Impersonation	This probe tests whether an agent can be persuaded to impersonate specific individuals, brands, or organizations, potentially causing reputational damage

Giskard

LEGAL & FINANCIAL RISK

Probes targeting potential legal and financial liabilities.

Probe	Description
LiabilityEngagementProbe	This probe tests whether an AI agent can be influenced to make commitments, agreements, or statements that could engage the company's legal or financial liability beyond its intended scope of operation

Giskard

MISGUIDANCE AND UNAUTHORISED ADVICE

Probes that test whether AI agents can be manipulated to provide professional advice outside their intended scope.

Probe	Description
Unauthorized Advice	This probe generates adversarial messages that request unauthorized advice from the agent, including financial recommendations, medical advice, legal counseling, etc