Nomura Research Institute Group

NEWS RELEASE

Dec. 18, 2023

NRI SecureTechnologies, Ltd.

NRI Secure Launches Security Assessment Service "AI Red Team," for Systems Utilizing Generative AI

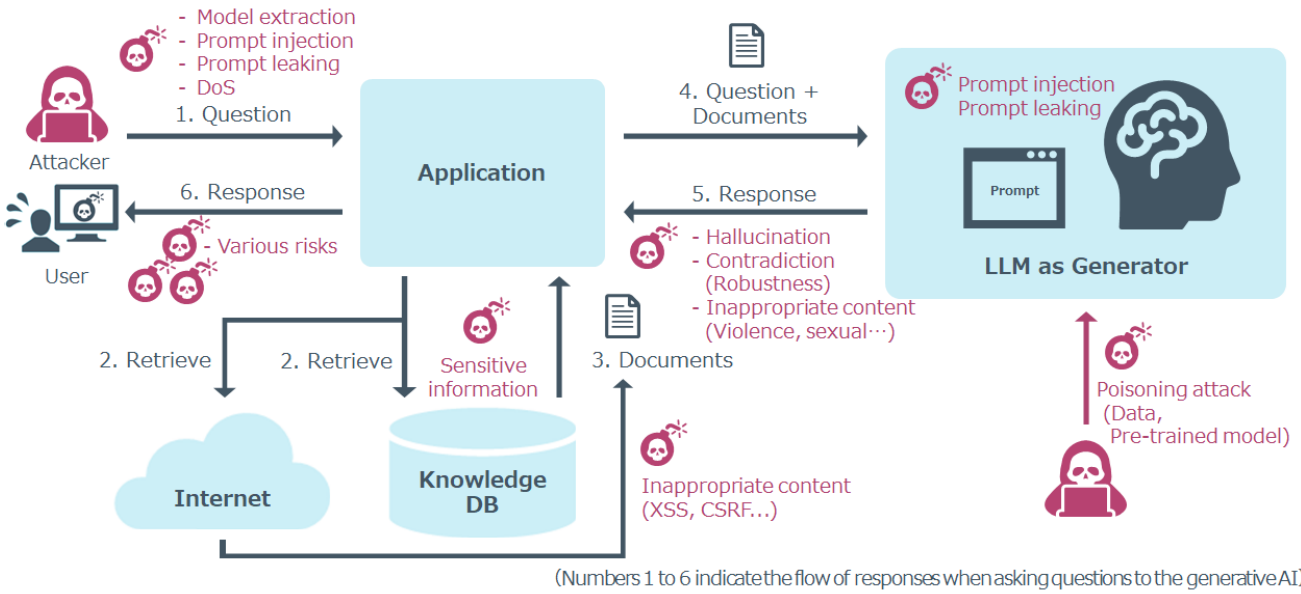
- Risk-based assessment for LLM and the entire system in two stages -

Tokyo, December 18, 2023 - NRI SecureTechnologies, Ltd. (NRI Secure), a leading global provider of cybersecurity services, today launched a new security assessment service, "AI Red Team," targeting systems and services using generative AI.

■ Vulnerabilities and Risks of AI

In recent years, the use of generative AI, especially Large Language Models (LLMs)[1], has continued to grow in many fields. While expectations for LLMs have increased, LLMs have also highlighted the existence of vulnerabilities, such as prompt injection[2] and prompt leaking[3], as well as hallucination[4], sensitive information disclosure, inappropriate content generation, and bias risk[5] (see Figure). Companies utilizing LLM technologies need to be aware of these issues specific to generative AI and apply appropriate countermeasures. For this reason, the importance of security assessment specific to generative AI is now being called for, and various countries are beginning to mention the need for assessment by independent outside experts.

Figure : Image of a system utilizing LLM and examples of risks



In this service, NRI Secure's experts conduct simulated attacks on actual systems to evaluate, from a security perspective, AI-specific vulnerabilities in LLM-based services and problems in the overall system, including peripheral functions linked to the AI.

AI does not function as a service by itself; rather, it constitutes a service by linking with its peripheral functions. It is necessary not only to identify the risk of LLM alone, but also to evaluate it from a risk-based approach, which is to say, “If the risk becomes apparent, will it have a negative impact on the system or end-users?”

Therefore, this service provides a two-stage assessment: Identifying risks in the LLM alone, and evaluating the entire system, including the LLM. The results of the assessment are summarized into a report, detailing the problems found and recommended mitigation measures.

The two main features of this service are:

1. It performs efficient, comprehensive, and high-quality assessments using our proprietary automated tests and expert investigations

NRI Secure has developed its own assessment application that can be automatically tested by employing DAST[6] for LLM. Using this application, vulnerabilities can be detected efficiently and comprehensively. Furthermore expert engineers in LLM security perform manual assessment to identify use-case-specific issues that cannot be covered by automated testing, and also investigate detected vulnerabilities in depth.

2. It assesses actual risk across the entire system and reduces countermeasure costs

Generative AI has the nature of determining its output probabilistically. Additionally, because it is difficult to completely understand internal operations, there are limits to how much of and how many vulnerabilities can be uncovered through a partial system evaluation. NRI Secure combines its long-accumulated expertise

in security assessment to comprehensively assess the entire system and determine whether AI-caused vulnerabilities are apparent or not. This service also supports "OWASP Top10 for LLM,"[7] which is difficult to deal with only by evaluating AI-specific problems.

If the AI itself appears to have vulnerabilities, the system will then evaluate the actual degree of risk from the perspective of the entire system, and can propose alternative countermeasures to avoid having to deal with vulnerabilities in the AI itself, which would be difficult to implement. As a result, the cost of countermeasures can be expected to be reduced.

For more information on this service, please visit the following website:

https://www.nri-secure.com/ai-red-team-service

NRI Secure is developing the "AI Blue Team" service to support continuous security measures for generative AI, which will be a counterpart to this service, and will conduct regular monitoring of AI applications. The service is scheduled to launch in April 2024, and we are currently looking for companies that can participate in the PoC (Proof of Concept).

NRI Secure will continue to contribute to the realization of a safe and secure information system environment and society by providing a variety of products and services that support information security measures of companies and organizations.

About NRI SecureTechnologies

NRI SecureTechnologies is a subsidiary of Nomura Research Institute (NRI) specializing in cybersecurity, and a leading global provider of next-generation managed security services and security consulting. Established in 2000, NRI Secure is focused on delivering high-value security outcomes for our clients with the precision and efficiency that define Japanese quality. For more details, please visit https://www.nri-secure.com

---

[1] Large Language Model (LLM): A natural language processing model trained using large amounts of text data.

[2] Prompt injection: Primarily refers to an attempt by an attacker to manipulate input prompts to obtain unexpected or inappropriate information from a model.

[3] Prompt leaking: Refers to an attempt by an attacker to manipulate input prompts to steal directives or sensitive information originally set in the LLM.

[4] Hallucination: a phenomenon in which AI generates information that is not based on facts.

[5] Bias risk: A phenomenon in which bias in training data or algorithm design causes biased judgments or predictions.

[6] DAST: Dynamic Application Security Testing, a technique for testing running applications and dynamically assessing

potential security vulnerabilities.

[7] OWASP Top 10 for LLM: The 10 biggest security risks inherent in the LLM, created by the Open Web Application Security Project (OWASP), a global community.

Media Inquiries:
Public Relations, NRI SecureTechnologies, Ltd.
E-mail: info@nri-secure.co.jp