Microsoft ha lanzado una plataforma de automatización de acceso abierto denominada PyRIT (acrónimo de Python Risk Identification Tool) diseñada para detectar proactivamente riesgos en sistemas de inteligencia artificial generativa (IA).

La herramienta de evaluación del equipo de ataque está concebida para *«permitir que todas las organizaciones a nivel global innoven de manera responsable con los últimos avances en inteligencia artificial»,* según Ram Shankar Siva Kumar, líder del equipo de ataque de IA en Microsoft.

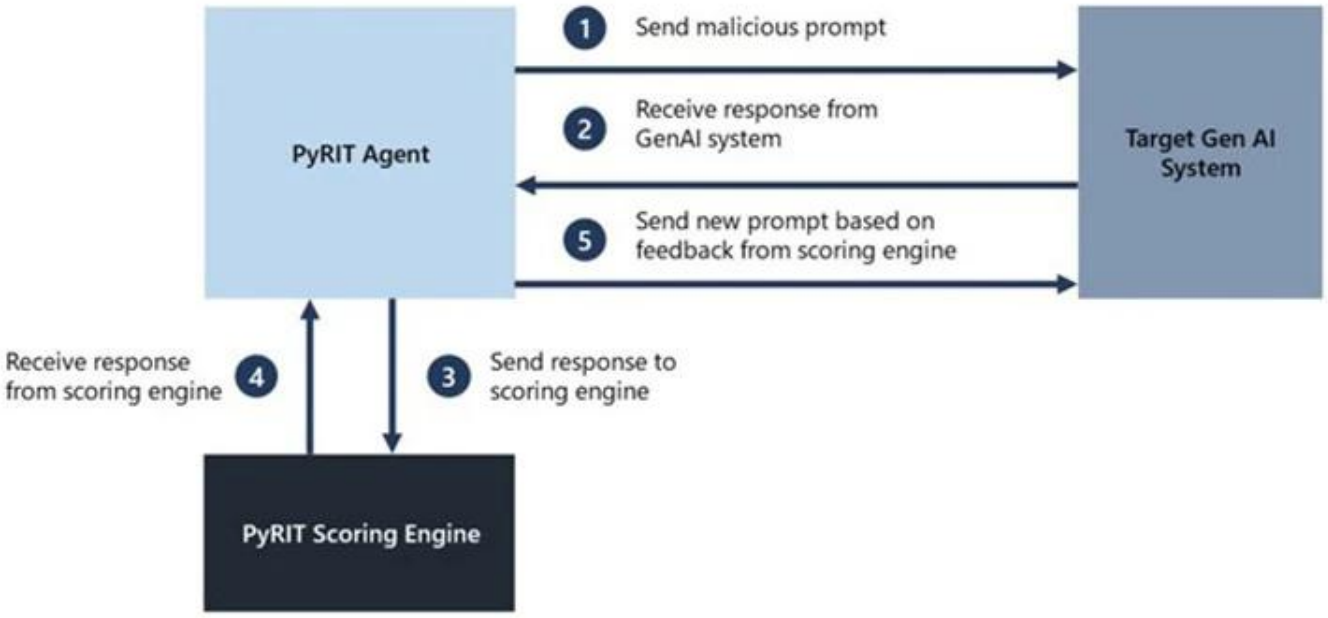
Según la compañía, PyRIT puede emplearse para evaluar la resistencia de los puntos finales de modelos de lenguaje extenso (LLM) frente a diversas categorías de riesgos, como la fabricación (por ejemplo, alucinaciones), el mal uso (por ejemplo, sesgo) y contenido prohibido (por ejemplo, acoso).

Además, la herramienta puede identificar daños en la seguridad que van desde la generación de malware hasta el jailbreaking, así como perjuicios a la privacidad como el robo de identidad.

PyRIT ofrece cinco interfaces: objetivo, conjuntos de datos, motor de puntuación, la capacidad de admitir múltiples estrategias de ataque e incorpora un componente de memoria que puede adoptar la forma de JSON o una base de datos para almacenar las interacciones de entrada y salida intermedias.

El motor de puntuación también proporciona dos opciones distintas para evaluar las salidas del sistema de IA objetivo, permitiendo a los equipos de ataque utilizar un clasificador de aprendizaje automático clásico o aprovechar un punto final LLM para la autoevaluación.

> *«El objetivo es permitir que los investigadores tengan una referencia de qué tan bien está desempeñándose su modelo y toda la tubería de inferencia frente a diferentes categorías de riesgo y poder comparar esa referencia con futuras iteraciones de su modelo»*, indicó Microsoft.



> *«Esto les permite contar con datos empíricos sobre el rendimiento actual de su modelo y detectar cualquier degradación basada en futuras mejoras».*

No obstante, la compañía enfatiza que PyRIT no reemplaza la evaluación manual de sistemas de IA generativos por parte de equipos de ataque y que complementa la experiencia existente del equipo en el dominio.

En otras palabras, la herramienta busca resaltar los «*puntos calientes*» de riesgo generando indicaciones que podrían utilizarse para evaluar el sistema de IA y señalar áreas que requieran una investigación más profunda.

Microsoft también reconoce que la evaluación de sistemas generativos de IA mediante equipos de ataque implica explorar simultáneamente riesgos de seguridad y de IA responsable, y que el ejercicio es más probabilístico, a la vez que destaca las notables diferencias en las arquitecturas de sistemas generativos de IA.

> *«La exploración manual, aunque lleva tiempo, a menudo es necesaria para identificar posibles áreas no cubiertas. La automatización es esencial para escalar, pero no reemplaza la exploración manual»,* afirmó Siva Kumar.

Este desarrollo se produce mientras Protect AI revela múltiples vulnerabilidades críticas en plataformas populares de la cadena de suministro de IA, como ClearML, Hugging Face, MLflow y Triton Inference Server, que podrían dar lugar a la ejecución arbitraria de código y la divulgación de información sensible.