Se ha dado a conocer una vulnerabilidad crítica de seguridad en SGLang que, si es explotada con éxito, podría permitir la ejecución remota de código en sistemas afectados.

La falla, identificada como CVE-2026-5760, cuenta con una puntuación CVSS de 9.8 sobre 10.0. Ha sido catalogada como un caso de inyección de comandos que deriva en la ejecución de código arbitrario.

SGLang es un framework de alto rendimiento y de código abierto diseñado para servir modelos de lenguaje de gran escala y modelos multimodales. Su repositorio oficial en GitHub ha sido bifurcado más de 5,500 veces y ha recibido 26,100 estrellas.

De acuerdo con el CERT Coordination Center (CERT/CC), la vulnerabilidad afecta al endpoint de reranking «/v1/rerank», permitiendo a un atacante lograr la ejecución de código arbitrario dentro del contexto del servicio SGLang mediante un archivo de modelo GPT-Generated Unified Format (GGUF) especialmente manipulado.

«Un atacante aprovecha esta vulnerabilidad creando un archivo de modelo GPT Generated Unified Format (GGUF) malicioso con un parámetro tokenizer.chat_template diseñado específicamente, que contiene una carga útil de inyección de plantillas del lado del servidor (SSTI) en Jinja2 junto con una frase de activación que desencadena la ruta de código vulnerable,» indicó CERT/CC en un aviso publicado hoy.

«Posteriormente, la víctima descarga y carga el modelo en SGLang, y cuando una solicitud alcanza el endpoint «/v1/rerank», la plantilla maliciosa se procesa, ejecutando código Python arbitrario del atacante en el servidor. Esta secuencia de eventos permite al atacante conseguir ejecución remota de código (RCE) en el servidor SGLang.»

Según el investigador de seguridad Stuart Beck, quien descubrió y reportó la vulnerabilidad, el problema raíz proviene del uso de jinja2.Environment() sin aislamiento (sandboxing), en lugar de emplear ImmutableSandboxedEnvironment. Esto posibilita que un modelo malicioso ejecute código Python arbitrario en el servidor de inferencia.

La secuencia completa del ataque es la siguiente:

- Un atacante crea un archivo de modelo GGUF con un tokenizer.chat_template malicioso que incluye una carga SSTI de Jinja2
- La plantilla incorpora la frase de activación del reranker Qwen3 para ejecutar la ruta de código vulnerable en «entrypoints/openai/serving_rerank.py»
- La víctima descarga y carga el modelo en SGLang desde fuentes como Hugging Face
- Cuando una solicitud llega al endpoint «/v1/rerank», SGLang lee el chat_template y lo renderiza usando jinja2.Environment()
- La carga SSTI ejecuta código Python arbitrario en el servidor

Cabe destacar que CVE-2026-5760 pertenece a la misma clase de vulnerabilidades que CVE-2024-34359 (también conocida como Llama Drama, con puntuación CVSS de 9.7), una falla crítica ya corregida en el paquete Python llama_cpp_python que podía permitir la ejecución de código arbitrario. Esta misma superficie de ataque también fue solucionada en vLLM a finales del año pasado (CVE-2025-61620, CVSS score: 6.5).

«Para mitigar esta vulnerabilidad, se recomienda utilizar ImmutableSandboxedEnvironment en lugar de jinja2.Environment() para procesar las plantillas de chat,» señaló CERT/CC. *«Esto evitará la ejecución de código Python arbitrario en el servidor. No se obtuvo respuesta ni parche durante el proceso de coordinación.»*