

# Laboratorios *Web LLM Attacks* PortSwigger Web Security Academy

14 de diciembre de 2025

Ibai Ruiz de Austri Lamas

En los laboratorios correspondientes al apartado de ataques a LLM se han explorado algunas de las vulnerabilidades que poseen estos sistemas. Entre ellas están el que la LLM tenga demasiados permisos, la inyección indirecta de prompts o el tratamiento inseguro de datos por parte de la propia LLM.

Se han realizado los laboratorios web de PortSwigger bajo la categoría de *Web LLM Attacks*, dentro de esta categoría hay 4 laboratorios online. El primero de estos explora el escenario en el que la LLM tiene demasiados permisos, permitiendo la ejecución de comandos SQL sobre la base de datos a petición de un usuario sin necesidad de credenciales.

En el segundo laboratorio se explora la vulnerabilidad a inyección de comandos en la detección de argumentos de la LLM, la LLM detecta la función a emplear en cada caso y los argumentos correspondientes, pero sin la correcta verificación de la validez de los argumentos enviados por el usuario es posible la ejecución remota de comandos escritos en la prompt enviada por el usuario.

En el tercer laboratorio se explora la inyección indirecta de prompts, esta vulnerabilidad ocurre debido a que la LLM no identifica correctamente cuando empieza o termina una prompt legítima de los datos que está tratando. Lo que supone que la LLM pueda llegar a ejecutar prompts no solicitadas por el usuario sino ocultas por el atacante entre la información manejada. El cuarto laboratorio explora una vulnerabilidad similar, excepto que en lugar de inyectar prompts se ejecutan scripts mediante la LLM debido a la falta de seguridad con respecto a Cross Site Scripting.

La LLM fue capaz de detectar en este último laboratorio la manipulación de la entrada para forzar acciones no deseadas cuando estas eran introducidas de manera directa en la información a tratar. La vulnerabilidad radicaba en que resulta posible ocultar el payload en el resto de la información tratada (en ese caso, una review de un producto) al darle un contexto razonable para estar ahí. Este contexto añadido puede llegar a evadir los filtros impuestos en LLMs vulnerables, que es lo que se explora en el laboratorio. En los casos en los que la LLM es capaz de identificar el payload malicioso este no se llega a mostrar al usuario ni a ejecutar, la efectividad de estos filtros internos en las LLM supone entonces un punto crítico a trabajar cuando se habla de seguridad debido a lo sensibles que estos sistemas son a manipulación externa, además de que prevalencia de LLMs en prácticamente todos los entornos digitales está en constante

aumento, siendo la seguridad de estos sistemas un tema de especial interés para el porvenir.