

Ciberseguridad Ofensiva para LLMs y Agentes

Taller • 14 horas • Zoom

Este taller te prepara para identificar, evaluar y mitigar vulnerabilidades en sistemas contruidos sobre modelos de lenguaje de gran tamaño (LLMs) y arquitecturas agénticas. El enfoque es práctico y ofensivo: mapearás la superficie de ataque probabilística de estos sistemas, explotarás integraciones complejas (RAG, tool calling, servidores MCP y memoria vectorial) y automatizarás pruebas adversariales (red teaming) mediante pipelines dedicados. Trabajarás sobre **casos de estudio reales**:

- **Caso EchoLeak**: Fuga de contexto y datos confidenciales mediante inyección indirecta de prompts en sistemas de memoria RAG indebidamente aislados.
- **Abuso de Herramientas vía MCP**: Escalamiento de privilegios desde un payload pasivo hacia la ejecución remota de comandos en el servidor (RCE).
- **Agentes Durmientes y Envenenamiento de Memoria**: Inyección de instrucciones maliciosas en bases de datos vectoriales que permanecen latentes hasta ser activadas por interacciones legítimas.

Todas las prácticas se ejecutan en entornos de laboratorio controlados y con fines defensivos: al terminar no solo sabrás cómo se rompen estos sistemas, sino cómo contenerlos.

Lo que aprenderás:

- Mapear la superficie de ataque probabilística de los LLM y entender por qué los esquemas de seguridad deterministas no bastan (OWASP LLM Top 10 y MITRE ATLAS).
- Evadir guardarraíles reales (LiteLLM con Llama Guard) mediante codificación Base64, ofuscación y secuencias de interacción multipaso.
- Explotar la inyección indirecta de prompts (IPI) y el envenenamiento de RAG para exfiltrar información mediante renderizado Markdown (OOB).
- Secuestrar el razonamiento de un agente (goal hijacking) y abusar del tool calling en grafos LangGraph, con observabilidad en Langfuse.
- Escalar privilegios a través del Model Context Protocol (MCP) hasta la ejecución remota de comandos (RCE) en contenedores Docker.
- Inyectar cargas latentes en memoria vectorial (ChromaDB) y analizar su propagación en sistemas multiagente.
- Automatizar el red teaming con Garak y PyRIT, midiendo el Attack Success Rate contra proxies de inferencia.
- Resolver un CTF integrador ("Breach the Agent") sobre una infraestructura completa y redactar el reporte de hallazgos técnicos.

Contenido

Semana 1 · Fundamentos ofensivos e inyección indirecta

- Sesión 1 — Jailbreaking y evasión de guardarraíles: superar esquemas de seguridad deterministas mediante la manipulación cognitiva del modelo en lenguaje natural. Mapeo OWASP/MITRE ATLAS, técnicas de ofuscación (cifrado, codificación) y ataques de optimización de prefijos. Práctica: evasión manual de LiteLLM (con Llama Guard) mediante codificación Base64 y secuencias multipaso sobre un modelo ligero desplegado con Unsloth.
- Sesión 2 — IPI, envenenamiento de RAG y exfiltración: explotar vulnerabilidades derivadas de la ingesta de fuentes externas no confiables. Vectores de confianza en arquitecturas RAG y exfiltración OOB mediante renderizado Markdown. Práctica: ingesta de documentos manipulados (LangChain + ChromaDB sin aislamiento de contexto) para inducir al modelo a concatenar credenciales internas dentro de peticiones externas.

Semana 2 · Secuestro agéntico y ecosistema MCP

- Sesión 3 — Goal hijacking y abuso de tool calling: interceptar la toma de decisiones autónoma y desviar los flujos de ejecución. Diferencia entre agencia y autonomía; funcionamiento interno de la invocación de herramientas (tool calling y esquemas JSON). Práctica: inyección vía comunicaciones simuladas y monitoreo de la alteración del razonamiento en grafos LangGraph con observabilidad en Langfuse, hasta forzar funciones no autorizadas.
- Sesión 4 — Protocolo MCP y escalamiento a RCE: evaluar riesgos de integración a nivel de sistema operativo y de protocolos especializados. Arquitectura del Model Context Protocol, fallas de validación y escalamiento de privilegios. Práctica: interceptación e inyección de comandos sobre los parámetros de servidores MCP locales en un contenedor Docker (python:slim) con validación flexible.

Semana 3 · Persistencia y automatización de pruebas

- Sesión 5 — Memoria persistente y comportamientos latentes: ataques asíncronos y contaminación de estado en almacenamiento vectorial. Memoria episódica frente a semántica y propagación de fallos en entornos multiagente. Práctica: inyección de cargas latentes condicionadas (LangChain + ChromaDB como memoria a largo plazo) para su detonación posterior ante consultas de terceros.
- Sesión 6 — Automatización de auditorías y evaluación adversarial: pasar de la explotación manual a esquemas estandarizados de prueba. Métricas de evaluación ofensiva (Attack Success Rate) y los frameworks Garak y PyRIT. Práctica: ejecución de pruebas automatizadas masivas contra proxies de inferencia y análisis de los reportes de vulnerabilidades generados.

Semana 4 · Ejercicio integrador

- Sesión 7 — Capstone CTF "Breach the Agent": aplicar las TTPs estudiadas en un entorno simulado de alta fidelidad, con infraestructura completa desplegada en contenedores (LiteLLM, LangGraph, ChromaDB y servidores MCP). Escenarios de extracción de información protegida, desvío de flujos mediante entradas no confiables y validación de controles de contención. Cierre: contramedidas, aislamiento (sandboxing) y validación estricta de esquemas, con debriefing y reporte de hallazgos técnicos sobre arquitecturas defensivas.

Al terminar tendrás:

- Un entorno de laboratorio ofensivo reproducible (LiteLLM, LangGraph, ChromaDB, servidores MCP y modelos locales con Unsloth) configurado y operado por ti durante las prácticas.
- La capacidad de ejecutar y explicar ataques de jailbreaking, inyección indirecta de prompts y envenenamiento de RAG, y de exfiltrar información en condiciones controladas.
- Experiencia práctica en secuestro de agentes (goal hijacking), abuso de tool calling y escalamiento de privilegios vía MCP hasta RCE.
- Pipelines de red teaming automatizado con Garak y PyRIT, con métricas de Attack Success Rate y reportes de vulnerabilidades interpretables.
- El criterio para mapear hallazgos a OWASP LLM Top 10 y MITRE ATLAS y proponer contramedidas: aislamiento (sandboxing), validación estricta de esquemas y guardarraíles.
- El proyecto integrador CTF "Breach the Agent" resuelto, con un reporte técnico de hallazgos.
- Constancia de participación verificable en línea.

Ya viste lo que vas a auditar.

Asegura tu lugar

Inscríbete ahora · \$1,400 MXN →

Constancia con registro STPS · Grabaciones por 1 año · Acceso en línea al instante

[¿Dudas antes de inscribirte? Escríbenos por WhatsApp](#)

Enfoque Ético y Descargo de Responsabilidad

Las tácticas, técnicas y procedimientos (TTPs) que se abordan en este taller se enseñan exclusivamente con fines académicos, éticos y defensivos: el objetivo es comprender cómo se vulneran estos sistemas para poder protegerlos mejor. Todas las prácticas ofensivas se ejecutan en entornos de laboratorio controlados, en aplicaciones deliberadamente vulnerables o en sistemas propios dispuestos para ese fin, nunca contra infraestructura de terceros. **Al inscribirse**, el participante se compromete a aplicar estos conocimientos únicamente sobre sistemas de su propiedad o sobre aquellos para los que cuente con autorización legal explícita y por escrito; emplearlos fuera de estos supuestos puede constituir una falta o un delito conforme a la legislación aplicable. **El participante es el único responsable del uso que haga de lo aprendido**: ni Actumlogos ni sus instructores se hacen responsables del mal uso de estos conocimientos ni de los daños o actos ilícitos que un estudiante pudiera cometer con ellos.

¿Para quién es este taller?

- Desarrolladores, ingenieros de datos y especialistas en TI que integran LLMs y agentes y necesitan asegurarlos.
- Profesionales de ciberseguridad que quieren especializarse en el frente ofensivo de la IA: red teaming de LLMs y sistemas agénticos.
- Ingenieros y arquitectos de IA que ya construyen con RAG, tool calling y MCP y quieren entender y cerrar sus vectores de ataque.

Este taller NO es para ti si:

- Buscas una introducción a Python, a los LLMs o a la ciberseguridad: aquí se asume base técnica previa.
- Quieres recetas de ataque para usarlas fuera de un entorno autorizado. Todo se practica en laboratorios controlados y con fines defensivos.
- Esperas un curso puramente teórico: el enfoque es práctico, ofensivo y basado en laboratorios.

Requisitos

Conocimientos fundamentales sobre la arquitectura y el funcionamiento de los LLMs; experiencia básica implementando e integrando proyectos de IA; manejo de Python a nivel intermedio y familiaridad con la línea de comandos (terminal); y conceptos básicos de ciberseguridad o un marcado interés en la especialización ofensiva. Se recomienda una computadora con Docker y, de preferencia, una GPU —propia o en la nube (gratis: Colab, Kaggle, LightningAI)— para ejecutar los modelos locales y seguir los laboratorios en vivo.

Stack Tecnológico (100% Código Abierto)

- **Hugging Face** y **Unsloth** para inferencia rápida y local.
- **LiteLLM** como proxy para estandarización de peticiones y guardarraíles de seguridad.
- **LangGraph** para flujos agénticos y **Langfuse** para observabilidad visual.
- **ChromaDB** para pruebas de envenenamiento de contexto en Python.
- Servidores Model Context Protocol (MCP) con **Postman** y scripts para inyección de payloads.
- **Garak** y **PyRIT** para escaneo y evaluación automatizada de vulnerabilidades.
- **Jupyter Notebooks** y contenedores **Docker** ligeros para la ejecución segura de pruebas.

Clases

4 semanas, 7 sesiones de 2 horas (14 h en total): lunes y jueves de 20:00 a 22:00 (hora de la Ciudad de México); fecha de inicio **5 de noviembre del 2026**. Se imparte en vivo en línea y todo queda grabado; tendrás acceso a las grabaciones durante 1 año para repasar a tu ritmo.

Instructor

M. en C. Alexis Terrazas — [LinkedIn](#)

Testimonios

Hemos capacitado a más de 2,000 personas en IA — visita actumlogos.com

Medios de pago

Tarjetas de crédito/débito, Link, PayPal, transferencia SPEI, Mercado Pago, Apple Pay, Google Pay. Ver [checkout](#)

Lista de Espera

Constancia de participación con folio único, verificable en línea

Al terminar recibirás una constancia de participación firmada por el instructor y el director de Actumlogos — agente capacitador registrado ante la Secretaría del Trabajo y Previsión Social. Cada constancia incluye un folio único con el que cualquier persona puede verificar su autenticidad en línea, y podrás compartirla directamente en LinkedIn.

Registro STPS: ZAGE-810930-FW2-0005

Este es un ejemplo para ilustrar qué esperar:

