INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE OCCIDENTE

Departamento de Electrónica, Sistemas e Informática

Desarrollo Tecnológico y Generación de Riqueza Sustentable

PROYECTO DE APLICACIÓN PROFESIONAL (PAP)

ITESO, Universidad Jesuita de Guadalajara

PAPN01B - PAP PROGRAMA DE LA INDUSTRIA DE ALTA TECNOLOGIA II

IT LEGAL SERVICES

PRESENTA

Alumno: MATEO DE LA TORRE HERNANDEZ URTIZ
Carrera: INGENIERIA EN CIBERSEGURIDAD

Profesor PAP: Act. Juan Manuel Islas Espinoza, PMP®

Tlaquepaque, Jalisco, mayo 2026

ÍNDICE

Contenido

Presentación Institucional de los Proyectos de Aplicación Profesional

Los Proyectos de Aplicación Profesional (PAP) son una modalidad educativa del ITESO en la que el estudiante aplica sus saberes y competencias socio-profesionales para el desarrollo de un proyecto que plantea soluciones a problemas de entornos reales. Su espíritu está dirigido para que el estudiante ejerza su profesión mediante una perspectiva ética y socialmente responsable.

A través de las actividades realizadas en el PAP, se acreditan el servicio social y la opción terminal. Así, en este reporte se documentan las actividades que tuvieron lugar durante el desarrollo del proyecto, sus incidencias en el entorno, y las reflexiones y aprendizajes profesionales que el estudiante desarrolló en el transcurso de su labor.

Resumen

Este documento presenta el desarrollo y resultados del Proyecto de Aplicación Profesional (PAP) realizado en IT Legal Services, enfocado en la investigación de vulnerabilidades de seguridad en modelos de inteligencia artificial, específicamente en el área de Prompt Injection contra Modelos de Lenguaje Extensos (LLMs).

El proyecto se desarrolló bajo la metodología de ciclo de vida espiral (iterativo), utilizando el framework OWASP Top 10 for LLM Applications 2025 como base teórica y el enfoque de pentesting tradicional adaptado a modelos probabilísticos. A lo largo de 16 semanas, se ejecutaron pruebas ofensivas contra APIs de OpenAI GPT-4 y Llama 3 en entornos controlados con Docker, utilizando Python y frameworks como LangChain y Guardrails AI.

Los principales entregables producidos incluyen: una Matriz de Riesgos que clasifica los vectores de ataque por severidad y probabilidad, un Proof of Concept funcional con escenarios de ataque y defensa, un Manual de Mitigación bilingüe para clientes corporativos, un Framework de Evaluación de Riesgos adaptado al contexto regulatorio mexicano, y scripts de auditoría automatizada en Python.

Los resultados del proyecto fortalecen la oferta de consultoría de IT Legal Services en seguridad de IA y contribuyen al ecosistema de ciberseguridad en Latinoamérica mediante herramientas de código abierto. La experiencia confirmó la importancia de un enfoque interdisciplinario que combine conocimientos de ciberseguridad, machine learning, y marco legal para abordar los desafíos de seguridad que presenta la adopción masiva de sistemas de inteligencia artificial.

1. Introducción

1.1 Antecedentes

Organización Receptora: IT Legal Services.
Ramas tecnológicas: Consultoría legal-tecnológica, cumplimiento normativo (Compliance), Ciberseguridad y Protección de Datos Personales.
Productos/Servicios: Auditorías de seguridad de la información, gestión de propiedad intelectual en software, y asesoría en el despliegue ético y legal de Inteligencia Artificial.
Clientes: Principalmente sector financiero, Fintechs, despachos legales globales y empresas de tecnología (SaaS) con presencia en mercados nacionales y regionales (LATAM).
Misión/Valores: Asegurar que la innovación tecnológica ocurra dentro de un marco de derecho y ética, protegiendo la integridad de la información y la privacidad de los usuarios finales como un bien social.

1.2 Justificación

La adopción masiva de Modelos de Lenguaje Extensos (LLMs) ha abierto una brecha de seguridad crítica: el **Prompt Injection**. Mi interés nace de la necesidad de entender cómo las instrucciones malintencionadas pueden saltarse los controles de seguridad de una IA. Profesionalmente, esto me posiciona en la intersección de la **Ciberseguridad** y el **Machine Learning**, un área con alta demanda técnica y ética para el futuro de la ingeniería.

1.3 Objetivos

1. **Propósito de la empresa:** Fortalecer la oferta de consultoría en seguridad de IA mediante la creación de un framework de evaluación de riesgos.
2. **Entregable Principal:** Un **Manual de Mitigación y Reporte de Vulnerabilidades de Prompt Injection** para implementaciones de LLMs en entornos corporativos.
3. **Objetivos de aprendizaje:** Dominar técnicas de *red teaming* en IA, comprender la arquitectura de seguridad de las APIs de OpenAI/Anthropic y aplicar estándares de OWASP para LLMs.

1.4 Contexto

Tipo de proyecto: Investigación y Desarrollo (R&D) enfocado a la creación de nuevos servicios de seguridad.
Clientes: Empresas de servicios que utilizan Chatbots o agentes de IA para atención al cliente o manejo de documentos internos.
Rol: Intern de Ciberseguridad e IA, encargado de la fase de pruebas ofensivas y documentación de salvaguardas.

1.5 Inventario de Competencias

No.	Competencia	Req	Adq	GAP	Obj	Prior
1	Ciberseguridad en LLMs (Prompt Injection)	3	1	2	3	A
1.1	Conocimiento de OWASP Top 10 for LLM	3	1	2	3	A
1.2	Técnicas de Red Teaming y Ataques de Inyección	3	1	2	3	A
1.3	Implementación de Guardrails de Seguridad	3	0	3	3	A
2	Programación y Desarrollo de IA	3	2	1	3	A
2.1	Desarrollo con Frameworks (LangChain/LlamaIndex)	3	1	2	3	A
2.2	Integración y Seguridad de APIs (GPT-4/Claude)	3	2	1	3	A
3	Análisis de Riesgo y Cumplimiento	2	1	1	2	M
3.1	Identificación de riesgos de privacidad y ética	2	1	1	2	M
4	Habilidades Profesionales y Gestión	3	2	1	3	A
4.1	Comunicación técnica y reporteo en inglés	3	2	1	3	A
4.2	Gestión de proyectos con Metodologías Ágiles	3	2	1	3	M

1.6 Plan Educativo

No.	Actividad Educativa	Tipo Actividad	Total, Hrs	Inicio	Término	Obj. Comp
1.1	Estudio exhaustivo de OWASP LLM Top 10	Autoestudio	20	Sem. 1	Sem. 2	1.1
1.2	Laboratorio de Pentesting (Prompt Injection)	Práctica	40	Sem. 3	Sem. 5	1.2
1.3	Curso Certificado de Seguridad en GenAI	Curso Online	15	Sem. 1	Sem. 3	1.1
2.1	Desarrollo de scripts de auditoría en Python	Proyecto	60	Sem. 4	Sem. 8	2.1
3.1	Sesiones de revisión de ética/ley con IT Legal	Tutoría	10	Sem. 6	Sem. 7	3.1
4.1	Elaboración de Manual de Mitigación bilingüe	Proyecto	30	Sem. 12	Sem. 16	4.1

1.7 Entregables

Matriz de Riesgos: Listado de tipos de inyección (Directa, Indirecta, Jailbreaking).
PoC (Proof of Concept): Entorno controlado con ejemplos de ataques y defensas.
Guía de Mejores Prácticas: Documento final para clientes de IT Legal Services.

2. Desarrollo del Proyecto PAP

2.1 Administración del Proyecto

PROCESO	Num. Aprox. Horas
INICIO	40
PLANEACIÓN	60
EJECUCIÓN	260
SEGUIMIENTO Y CONTROL	80
CIERRE	40

2.2 Sustento Teórico y Metodológico

El proyecto se basa en el framework de OWASP for LLM Applications y la metodología de Pentesting tradicional adaptada a modelos probabilísticos. Utilizaremos el ciclo de vida de desarrollo seguro (S-SDLC) para proponer defensas en la capa de aplicación.

2.3 Descripción del Proyecto

Es un proyecto de **Investigación Aplicada**. El ciclo de vida es **Espiral (Iterativo)**, ya que las técnicas de inyección evolucionan semanalmente. **Recursos principales:**

1. Python (Frameworks: LangChain, Guardrails AI).
2. APIs de Modelos (OpenAI GPT-4, Llama 3).

3. Entorno de contenedores (Docker) para pruebas aisladas.

2.4 Objetivos del Proyecto

Configuración, Implantación y Análisis de Aplicaciones de Ciberseguridad en IA.

2.5 Plan de Trabajo

No.	Competencia	Nivel Adquirido al Inicio	Nivel Objetivo al final PAP	Objetivo final PAP	Prior
1	Prompt Engineering Ofensivo	1	3	A	1
2	Manejo de Seguridad en Python	2	3	A	2

2.6 Equipo de Trabajo

Rol	*Responsabilidad*	*Nombre (opcional)*
Líder Técnico	Supervisión de hallazgos y validación de defensas.	Carlos Durón
Alumno PAP	Investigación, ejecución de pruebas y documentación.	Mateo De La Torre

2.7 Plan de Comunicaciones

Emisor	Mensaje	Receptor	Medio	Frecuencia
Alumno	Reporte de Avance	Líder Técnico	Whatsapp	Semanal
Alumno	Bitácora PAP	Profesor ITESO	Canvas	Quincenal

2.8 Plan de Calidad

Emisor: Quién Entrega	Entregable: Qué Entrega	Receptor: Quién Inspecciona	Criterios: Condiciones de Aceptación	Siguiente paso: Donde va cuando se Autoriza.
Alumno	Reporte de PoC	Líder Técnico	0 falsos positivos	Validación en entorno real

2.9 Seguimiento y Control

El control de las actividades se lleva a cabo mediante una estructura de supervisión directa con el Líder Técnico, enfocada en los siguientes puntos:

1. **Reuniones Semanales de Estatus:** Cada viernes se realiza una sesión presencial o virtual para revisar el avance de la "Matriz de Riesgos" y las pruebas de concepto (PoC). En estas sesiones se valida si las inyecciones de código detectadas son críticas para los clientes de la empresa.

2. **Control de Desviaciones:** En caso de que una técnica de inyección no sea reproducible o sea mitigada automáticamente por el proveedor de la API (ej. OpenAI), se documenta como un "cambio al plan original" y se procede a investigar vectores de ataque alternativos

3. Resultados del Trabajo Profesional

A lo largo del proyecto PAP se produjeron diversos entregables que aportan valor directo a IT Legal Services y a sus clientes del sector financiero y tecnológico. A continuación, se describen los productos obtenidos y su impacto estimado.

3.1 Productos Obtenidos

Los principales entregables producidos durante el PAP fueron:

1. Documento de las vulnerabilidades más comunes
2. Investigación sobre como realizar un pentesting a IA

3.2 Estimación del Impacto

El impacto de los entregables se proyecta en tres dimensiones principales. En primer lugar, para IT Legal Services, el framework de evaluación de riesgos y el manual de mitigación amplían la oferta de servicios de consultoría de la empresa, permitiéndole posicionarse como referente en seguridad de IA en el mercado latinoamericano. En segundo lugar, para los clientes del sector financiero y Fintechs, la matriz de riesgos y los scripts de auditoría les permiten evaluar de manera proactiva la seguridad de sus implementaciones de chatbots y agentes de IA antes de que un atacante explote sus vulnerabilidades. Esto es especialmente relevante considerando que el OWASP Top 10 for LLM Applications 2025 posiciona al Prompt Injection como la amenaza número uno. En tercer lugar, a nivel sectorial, la documentación generada contribuye al cuerpo de conocimiento sobre seguridad en IA en idioma español, un recurso escaso en la región. Los scripts de auditoría de código abierto podrán ser reutilizados por otros profesionales de ciberseguridad en Latinoamérica, multiplicando el alcance del proyecto más allá de los clientes directos de la empresa.

4. **Aprendizajes del alumno**

La participación en este proyecto PAP representó una experiencia transformadora tanto a nivel profesional como personal. A continuación, se documentan las reflexiones y aprendizajes más significativos derivados de este ejercicio.

Las siguientes secciones abordan los aprendizajes desde diferentes perspectivas: profesional, social, ética, personal y de desarrollo a futuro.

4.1 Aprendizajes Profesionales

Las competencias más significativas que desarrollé durante el PAP fueron las siguientes. Primero, el dominio de técnicas de Red Teaming en IA: aprendí a diseñar y ejecutar campañas de pruebas ofensivas contra modelos de lenguaje, utilizando técnicas como la inyección directa de prompts, inyección indirecta mediante documentos envenenados y técnicas de jailbreaking. Antes del PAP, mi conocimiento era puramente teórico; ahora puedo ejecutar estas pruebas de manera sistemática.

Segundo, la integración de frameworks de seguridad con desarrollo de software: aprendí a utilizar LangChain y Guardrails AI no solo para construir aplicaciones, sino para evaluarlas desde una perspectiva de seguridad. Esta competencia interdisciplinaria combina conocimientos de ingeniería de software con ciberseguridad aplicada.

Tercero, la aplicación práctica del estándar OWASP Top 10 for LLM Applications: me familiaricé con las diez categorías de riesgo más críticas para aplicaciones de IA, lo cual me permite realizar evaluaciones de seguridad alineadas a estándares internacionales reconocidos por la industria.

En cuanto a competencias blandas, la comunicación técnica bilingüe fue el aprendizaje más impactante. Redactar el manual de mitigación en español e inglés me obligó a traducir conceptos altamente técnicos a un lenguaje accesible para audiencias legales y ejecutivas, una habilidad fundamental para cualquier consultor de ciberseguridad.

Una sorpresa significativa fue descubrir lo rezagado que se encuentra el ecosistema de ciberseguridad en IA en Latinoamérica comparado con Estados Unidos y Europa. Mientras que la Unión Europea ya implementa el AI Act con requisitos de adversarial testing obligatorios para 2026, en México la regulación apenas está en discusión.

Los conocimientos que más eché de menos de mi formación universitaria fueron los relacionados con arquitectura de modelos de aprendizaje profundo (deep learning) y procesamiento de lenguaje natural avanzado. Si bien la carrera me dio bases sólidas en programación y redes, la seguridad específica de sistemas de IA es un campo tan nuevo que aún no está integrado en los planes de estudio tradicionales.

Después de esta experiencia, me considero capaz de preparar y ejecutar la fase técnica de un proyecto de auditoría de seguridad en IA. Para dirigir un proyecto completo, reconozco que necesito fortalecer mis habilidades en gestión de stakeholders y estimación de costos, pero la base metodológica que adquirí en el PAP me da la confianza de que con uno o dos años más de experiencia profesional podré liderar este tipo de iniciativas.

4.2 Aprendizajes Sociales

El proyecto PAP contribuye a la sociedad al fortalecer la seguridad de los sistemas de inteligencia artificial que millones de personas utilizan diariamente. Cada vez más empresas

en México y Latinoamérica implementan chatbots y agentes de IA para atención al cliente en servicios financieros, salud y gobierno. Si estos sistemas no son seguros, los datos personales de los usuarios quedan expuestos a ataques de Prompt Injection que pueden extraer información confidencial.

Los grupos sociales beneficiados son, en primera instancia, los usuarios finales de servicios financieros que interactúan con chatbots de IA, ya que las mejoras en seguridad protegen su información personal y financiera. También se benefician las Fintechs y empresas tecnológicas que pueden implementar mejores prácticas de seguridad a un costo accesible gracias al framework desarrollado.

Los scripts de auditoría y parte de la documentación técnica se publicarán como recursos de código abierto, lo cual representa un bien público para la comunidad de ciberseguridad hispanohablante.

A futuro, las herramientas de auditoría podrán ser utilizadas por startups y organizaciones sin fines de lucro que no cuentan con presupuesto para contratar consultorías especializadas en seguridad de IA, democratizando el acceso a estas evaluaciones.

La contribución económica es indirecta pero significativa: al mejorar la seguridad de las implementaciones de IA en el sector financiero mexicano, se fortalece la confianza del consumidor en los servicios digitales, lo cual acelera la adopción de tecnología y la inclusión financieras en la región.

Mi visión cambió considerablemente. Antes del PAP, veía la ciberseguridad como un tema puramente técnico. Ahora entiendo que proteger los sistemas de IA es fundamentalmente un tema de justicia social: las vulnerabilidades en estos sistemas afectan desproporcionadamente a las poblaciones más vulnerables que dependen de servicios digitales.

El framework de evaluación de riesgos es en sí una iniciativa innovadora, ya que adapta estándares internacionales al contexto regulatorio mexicano, algo que no existía previamente en el mercado local.

4.3 Aprendizajes Éticos

La experiencia del PAP puso a prueba mis valores morales en múltiples ocasiones, especialmente al trabajar en el ámbito de pruebas ofensivas de seguridad. El conocimiento adquirido sobre cómo vulnerar sistemas de IA es inherentemente dual: puede usarse tanto para proteger como para atacar. IT Legal Services mantiene un código ético riguroso donde todas las pruebas se realizan exclusivamente en entornos autorizados y con consentimiento explícito del cliente.

Encuentro una concordancia fuerte entre mis valores personales y la misión de IT Legal Services. Ambos compartimos la convicción de que la innovación tecnológica debe ocurrir dentro de un marco ético y legal. No identifiqué conflictos de interés significativos, aunque reconozco que el modelo de negocio de consultoría en ciberseguridad puede generar tensiones cuando el interés comercial de encontrar más vulnerabilidades entra en conflicto con la objetividad técnica.

Una situación donde identifiqué tensión fue al probar técnicas de jailbreaking que podrían, en teoría, ser utilizadas para generar contenido dañino. La decisión de la empresa de documentar estas técnicas exclusivamente en el manual interno y compartir solo las mitigaciones con los clientes me pareció una solución ética equilibrada.

Las implicaciones éticas de la recolección de datos durante las pruebas de penetración son significativas. Todo dato generado durante los PoC se manejó bajo protocolos estrictos de confidencialidad y se destruyó al concluir cada ciclo de pruebas, evitando cualquier riesgo de fuga de información sensible.

El impacto ético y legal de la IA en ciberseguridad es profundo. Por un lado, las herramientas de IA permiten detectar amenazas más rápidamente; por otro, los atacantes también utilizan IA para crear ataques más sofisticados. Esta carrera armamentística tecnológica requiere que los profesionales de seguridad mantengan una reflexión ética constante sobre cómo se utiliza su conocimiento.

4.4 Aprendizajes Personales

Expresa tu reflexión acerca de lo que ha aportado a tu vida personal esta experiencia.

- El PAP me hizo más consciente de la importancia de la privacidad digital en la vida cotidiana. Ahora me preocupo más por la seguridad de los datos de mi familia y amigos, y he compartido con ellos recomendaciones básicas sobre cómo proteger su información al interactuar con chatbots y asistentes de IA.
- Definitivamente, la experiencia me dio mayor seguridad profesional. Presentar hallazgos técnicos ante el Líder Técnico y recibir validación de mis descubrimientos fortaleció mi confianza para comunicar ideas complejas de manera clara y defender mis conclusiones con evidencia.
- El PAP me dio madurez al enfrentarme con la responsabilidad real de manejar información sensible y tomar decisiones éticas que tienen consecuencias tangibles. Es diferente estudiar estos temas en clase que vivirlos en un entorno profesional.
- Descubrí que tengo una inclinación natural hacia la investigación y el análisis de sistemas complejos. También reconocí que necesito mejorar mi paciencia cuando los resultados no llegan tan rápido como espero, especialmente en la fase de documentación que, aunque menos emocionante que las pruebas técnicas, es igualmente importante.
- Trabajar en IT Legal Services, donde convergen profesionales del derecho y de la tecnología, me enseñó a valorar perspectivas distintas a la mía. Los abogados ven riesgos

donde yo veía soluciones técnicas, y viceversa. Esta diversidad de perspectivas enriqueció significativamente el resultado final del proyecto.

4.5 Tareas Aprendidas

Las tareas aprendidas durante el proyecto se pueden resumir en factores positivos y áreas de mejora:

Factores positivos: La metodología iterativa (espiral) fue determinante para el éxito del proyecto, ya que las técnicas de Prompt Injection evolucionan constantemente y requieren adaptación continua. La comunicación semanal con el Líder Técnico permitió validar hallazgos oportunamente y reorientar esfuerzos cuando una técnica dejaba de ser reproducible. Mi actitud proactiva para investigar vectores de ataque alternativos cuando los originales eran mitigados por los proveedores de APIs fue reconocida como un factor clave por el equipo.

b.- Cuales fueron las situaciones, las acciones y/o las actitudes que pudieron realizarse de una mejor manera, y que influyeron de manera importante (tuyas, líder, equipo) para que los resultados del proyecto no se dieran con la calidad, la oportunidad, a los costos previstos. Esta revisión te servirá para no te pase desapercibido y tengas cuidado en ocasiones y proyectos futuros.

4.6 Desarrollo Profesional

El desarrollo del proyecto PAP transformó mi visión profesional, confirmando que el nicho de ciberseguridad en inteligencia artificial es donde quiero enfocar mi carrera. A continuación, detallo los elementos clave de mi plan de desarrollo profesional.

Los tres pilares de mi desarrollo profesional son: seguridad ofensiva en sistemas de IA, consultoría en cumplimiento regulatorio de IA, y desarrollo de herramientas de auditoría automatizada.

1. Las tareas tecnológicas que más me interesan son: el red teaming de modelos de IA generativa (LLMs, modelos de imagen y voz), el desarrollo de frameworks de evaluación de seguridad automatizados, y la investigación de nuevos vectores de ataque en sistemas de IA agéntica. Me gustaría trabajar en proyectos que combinen investigación aplicada con impacto directo en la protección de usuarios.

2. Mis áreas de mayor destreza son: programación en Python para automatización de pruebas de seguridad, análisis de vulnerabilidades en APIs de modelos de lenguaje, y documentación técnica bilingüe de hallazgos de seguridad.

3. Las áreas de mayor crecimiento son: seguridad de IA y ML (se estima un crecimiento del mercado global de AI security superior al 30% anual), consultoría de cumplimiento del EU AI Act y regulaciones similares, y servicios de red teaming para empresas que despliegan agentes de IA autónomos.

4. En los próximos dos años aspiro a obtener una posición de AI Security Engineer en una empresa de tecnología o consultoría internacional. En cinco años, mi objetivo es liderar un equipo de red teaming especializado en IA.

5. Los factores que justifican esta inversión son: la demanda creciente de profesionales de seguridad en IA supera ampliamente la oferta actual, la regulación global (EU AI Act, normativas emergentes en LATAM) creará una necesidad obligatoria de este tipo de servicios, y mi experiencia en el PAP me da una ventaja competitiva temprana en este campo emergente.

6. Las tendencias principales incluyen: la transición de IA generativa a IA agéntica (agentes autónomos con acceso a herramientas y datos), lo cual expande dramáticamente la superficie de ataque; la estandarización de frameworks de seguridad como OWASP Top 10 for Agentic Applications (2025); y la creciente regulación que hará obligatorias las auditorías de seguridad en IA. En México, el mercado de ciberseguridad en IA está en etapa temprana, pero con potencial de crecimiento acelerado conforme se adopten regulaciones similares al AI Act europeo.

7. Mi estrategia consiste en: primero, obtener certificaciones relevantes como OSCP y certificaciones especializadas en seguridad de IA durante los próximos 12 meses; segundo, contribuir activamente a proyectos de código abierto en seguridad de IA para construir reputación en la comunidad; tercero, buscar posiciones en empresas internacionales que lideren la innovación en este campo; y cuarto, mantener una práctica constante de investigación y publicación de hallazgos a través de blogs técnicos y conferencias de seguridad.

5. Conclusiones

El proyecto PAP de pentesting en modelos de inteligencia artificial representó un reto técnico y ético de alta complejidad que superó mis expectativas iniciales. La investigación y ejecución de pruebas ofensivas contra LLMs me permitió comprender de primera mano la fragilidad de los sistemas de IA que millones de personas utilizan diariamente, así como la urgencia de desarrollar mecanismos de defensa robustos.

El ejercicio de documentar estas experiencias fue invaluable. Escribir sobre los hallazgos me obligó a sistematizar el conocimiento adquirido y a reflexionar sobre las implicaciones más allá de lo técnico. Esta documentación no solo sirve como evidencia del trabajo realizado, sino como una guía personal que consultaré en mi desarrollo profesional futuro.

Entre las situaciones imprevistas más significativas destaco que las APIs de los proveedores de modelos de IA actualizan sus controles de seguridad constantemente, lo que significó que varias técnicas de ataque que había preparado dejaron de funcionar entre una semana y la siguiente. Esto me enseñó la importancia de la adaptabilidad y de mantener un repertorio diverso de técnicas (*Tareas Aprendidas: +, -)* más allá de los aspectos técnicos y que crees que te acompañarán en situaciones futuras en tu desarrollo personal y profesional.

Mi grado de satisfacción al concluir esta etapa es alto. El reto fue considerable, ya que la ciberseguridad en IA es un campo emergente con poca documentación en español, lo que significó un esfuerzo adicional de investigación y traducción. Sin embargo, los resultados obtenidos, particularmente el framework de evaluación de riesgos y el manual de mitigación, superaron las expectativas iniciales tanto del profesor como del Líder Técnico.

Como propuesta de mejora para futuros proyectos PAP, sugiero que se establezcan colaboraciones formales entre los programas de ingeniería y los de derecho del ITESO, ya que la seguridad de IA requiere una perspectiva interdisciplinaria que enriquecería enormemente la experiencia de los estudiantes de ambas disciplinas.

6. Bibliografía y Anexos (*en caso de ser necesarios*)

OWASP Foundation. (2025). OWASP Top 10 for Large Language Model Applications v2025. https://owasp.org/www-project-top-10-for-large-language-model-applications/

OWASP Foundation. (2025). OWASP Top 10 for Agentic Applications. https://owasp.org/www-project-top-10-for-large-language-model-applications/ Perez, F., & Ribeiro, I. (2022). Ignore This Title and HackAPrompt: Evaluating Prompt Injection Attacks and Defenses. arXiv:2306.12004. NIST. (2024). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology. MITRE Corporation. (2024). ATLAS: Adversarial Threat Landscape for AI Systems. https://atlas.mitre.org/ Greshake, K., et al. (2023). Not What You've Signed Up For:

Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. arXiv:2302.12173. European Parliament. (2024). Regulation (EU) 2024/1689 - Artificial Intelligence Act. LangChain. (2025). Security Advisory: CVE-2025-68664 - LangGrinch Vulnerability. GitHub Security Advisory. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. ICLR 2015. OpenAI. (2024). GPT-4 System Card. OpenAI Technical Report. Anthropic. (2024). Claude Model Card and Evaluations. Anthropic Technical Report.