LARAF-T: Un Marco Forense Digital para el Análisis de Modelos de Inteligencia Artificial Personalizados y Operados por Terceros en Entornos de Ciberseguridad

Francisco René Lara Acute
Maestría Profesional en CiberSeguridad, Escuela de Ciberseguridad, Universidad Cenfotec, San José, Costa Rica 2026

***Resumen*—**La creciente adopción de modelos de Inteligencia Artificial (IA) operados por terceros bajo esquemas Software-as-a-Service (SaaS) en entornos corporativos de ciberseguridad ha generado un vacío crítico en las capacidades de análisis forense post-incidente. Las metodologías forenses digitales existentes no fueron diseñadas para abordar la opacidad y las restricciones de acceso propias de los sistemas de IA de caja negra. Este artículo presenta LARAF-T (Layered Response Attribution and Forensic Traceability), un marco forense estructurado en seis fases para la investigación post-incidente de modelos de IA personalizados operados por proveedores externos. El marco integra principios de ISO/IEC 27037:2012, NIST AI RMF, NIST AI 600-1 y OWASP LLM Top 10 (2024). La validación mediante cuatro escenarios de ataque controlados —inyección de prompts (E-01), envenenamiento de datos (E-02), divulgación de información sensible (E03) y ataque combinado multi-vector (E-04)— produjo tasas de detección del 80%, 100%, 93.3% y 80% respectivamente, bajo condiciones simuladas de acceso SaaS. LARAF-T contribuye una metodología replicable y tecnológicamente neutral que fortalece la cadena de custodia y el valor probatorio de los artefactos digitales generados en incidentes de seguridad relacionados con IA.

Palabras Clave*—forense de IA, forense digital, inyección de prompts, envenenamiento de datos, cadena de custodia, seguridad SaaS, OWASP LLM Top 10, NIST AI RMF, respuesta a incidentes, IA explicable, modelos de terceros, modelos de caja negra.*

I. Introducción

La rápida integración de servicios de Inteligencia Artificial en las operaciones de ciberseguridad corporativa ha transformado profundamente el panorama de la investigación post-incidente. Los sistemas basados en IA se encuentran actualmente embebidos en funciones críticas de seguridad: detección de intrusiones, análisis de comportamiento, respuesta automática a amenazas y evaluación de vulnerabilidades. Muchos de estos sistemas se adquieren como servicios gestionados o bajo modalidad SaaS de proveedores especializados, lo que significa que la organización consumidora interactúa con el modelo exclusivamente a través de APIs o interfaces web, sin visibilidad sobre la arquitectura subyacente, los datos de entrenamiento ni el pipeline de inferencia.

Esta opacidad arquitectural da lugar a lo que este trabajo denomina la *frontera de opacidad forense*: el punto a partir del cual una organización consumidora pierde visibilidad técnica sobre el estado interno del modelo. Según el informe Cost of a Data Breach de IBM (2025), el 29% de los incidentes de seguridad relacionados con IA se originan en modelos operados por terceros, superando tanto a los despliegues internos (26%) como a los de código abierto [1]. A pesar de esta prevalencia, la comunidad forense aún no ha producido una metodología estandarizada y validada para la investigación postincidente de tales sistemas.

Las metodologías forenses digitales tradicionales, codificadas en ISO/IEC 27037:2012, NIST SP 800-86 y NIST SP 800-61, presuponen el acceso del investigador a sistemas de archivos, memoria volátil, listados de procesos y registros de auditoría completos [2][3][12]. Estos supuestos son estructuralmente incompatibles con el modelo de IA SaaS de caja negra, donde la organización consumidora recibe únicamente logs de uso agregados de la API en respuesta a una solicitud de incidente.

Los marcos normativos actuales reconocen esta brecha sin resolverla. El OWASP Top 10 for LLM Applications (2024) identifica la inyección de prompts (LLM01) y la divulgación de información sensible (LLM06) como riesgos de primer nivel, pero su orientación se dirige hacia controles preventivos en la fase de desarrollo, no hacia el análisis forense post-incidente [4]. NIST AI 600-1 (2024) cataloga doce categorías de riesgo para sistemas de IA generativa sin ofrecer guía procedimental forense [5]. Una revisión sistemática de la literatura en IEEE Xplore y SpringerLink (2019–2025) confirmó la ausencia de marcos forenses estandarizados y validados que aborden específicamente el análisis post-incidente de modelos de IA de terceros. Este artículo aborda esa brecha proponiendo LARAF-T, un marco forense de siete fases diseñado para investigaciones post-incidente bajo condiciones de acceso restringido. Las contribuciones principales son: (i) un marco forense estructurado y tecnológicamente neutral que integra principios de forense digital con taxonomías de ataque específicas

de IA; (ii) una tipología de evidencia para sistemas de IA; (iii) un modelo de acceso estratificado que permite investigación productiva en escenarios de caja negra, cooperativos y de caja blanca; y (iv) validación empírica en cuatro escenarios de ataque controlados con tasas de detección cuantificadas.

II. Antecedentes y Trabajos Relacionados

A. Fundamentos de Forense Digital

La forense digital como disciplina ha abordado históricamente la adquisición y el análisis de datos provenientes de endpoints, infraestructura de red y medios de almacenamiento. El proceso se divide típicamente en cuatro fases —identificación, preservación, análisis y presentación— cada una regida por estándares diseñados para garantizar la integridad, autenticidad y admisibilidad legal de la evidencia [2]. Central a este proceso es la cadena de custodia: un registro documentado e ininterrumpido de quién manejó la evidencia, cuándo y cómo.

ISO/IEC 27037:2012 establece directrices para la identificación, recolección, adquisición y preservación de evidencia digital, enfatizando la primacía de la integridad de la evidencia mediante verificación criptográfica [2]. NIST SP 800-86 extiende esta guía a contextos de respuesta a incidentes, definiendo roles, herramientas y procedimientos para el manejo de evidencia en entornos empresariales [3]. RFC 3227 proporciona directrices para la recolección de evidencia electrónica ordenada por volatilidad, un concepto directamente aplicable a los artefactos de sistemas de IA, que pueden ser sobreescritos por actualizaciones del modelo o rotación de logs.

B. Vectores de Ataque Específicos de IA

La superficie de ataque de los modelos de IA se extiende más allá de las vulnerabilidades de software tradicionales. Varias clases de ataque son particularmente relevantes para los despliegues SaaS de IA de terceros:

- **Inyección de Prompts (OWASP LLM01):** Entradas adversariales diseñadas para anular instrucciones a nivel de sistema y alterar el comportamiento del modelo. En modelos SaaS de caja negra, el desafío forense consiste en detectar la inyección sin acceso al pipeline de procesamiento de entradas del modelo.
- **Envenenamiento de Datos:** Contaminación de conjuntos de datos de entrenamiento o ajuste fino para introducir sesgos, puertas traseras o degradación del rendimiento. La detección requiere análisis comparativo de versiones del modelo o líneas base de comportamiento.
- **Divulgación de Información Sensible (OWASP LLM06):** Extracción inadvertida o forzada de información confidencial incorporada en prompts del sistema o datos de entrenamiento. Este vector es especialmente peligroso en despliegues empresariales donde los modelos se inicializan con contexto operacional sensible.
- **Inversión y Extracción de Modelos:** Reconstrucción de datos de entrenamiento o parámetros del modelo mediante consultas sistemáticas. Detectable a través de patrones de consulta anómalos en los logs de la API.
- **Ejemplos Adversariales:** Entradas elaboradas que causan clasificación incorrecta. Para modelos de clasificación en aplicaciones de seguridad, esto puede derivar en evasión sistemática.

C. Marcos Existentes y Análisis de Brechas

Varios marcos abordan preocupaciones adyacentes pero no alcanzan a proporcionar una metodología forense completa para modelos de IA de terceros. La Tabla I resume el análisis de brechas.

Marco	Alcance	Cobertura Forense	Brecha
ISO/IEC 27037:2012	Evidencia digital tradicional	Recolección, preservación, cadena de custodia	Sin procedimientos para IA
NIST SP 800-86	Forense en respuesta a incidentes	Fases de investigación, roles	Sin forense de modelos IA
NIST AI RMF 1.0	Gobernanza de riesgo en IA	Identificación de riesgos, preparación	Sin procedimientos post-incidente
NIST AI 600-1	Riesgos de IA generativa	Taxonomía de riesgos, mitigaciones	Sin metodología forense
OWASP LLM Top 10	Seguridad de aplicaciones LLM	Taxonomía de vulnerabilidades	Solo enfoque preventivo
LARAF-T (este trabajo)	Forense post-incidente de modelos IA	Marco de siete fases, tipología de evidencia, CoC	— (solución propuesta)

Tabla I. Análisis de Brechas Forenses en Marcos Existentes.

D. Investigación Relacionada

La literatura académica sobre seguridad en IA ha crecido sustancialmente desde 2019, con contribuciones significativas en robustez adversarial [6]. Papernot et al. establecieron la taxonomía fundacional para amenazas de aprendizaje automático adversarial, que informa los vectores de ataque abordados por LARAF-T [6]. El trabajo en interpretabilidad (XAI), especialmente LIME [7] y SHAP [8], proporciona herramientas técnicas aplicables en la Fase F-4 del marco, aunque su despliegue presupone acceso de caja blanca o cooperativo. Según el conocimiento de los autores, ningún trabajo previo ha propuesto y validado empíricamente un marco forense completo orientado específicamente a la investigación de modelos de IA de terceros bajo restricciones de acceso de caja negra.

III. Modelo de Amenaza y Alcance

A. Contexto Operacional

LARAF-T apunta al siguiente escenario operacional: una organización (*consumidora*) ha desplegado un sistema de IA de ciberseguridad provisto por un tercero (*proveedor*) como servicio SaaS. La consumidora interactúa con el modelo a través de una API definida. Ocurre —o se sospecha— un incidente de seguridad que involucra la posible manipulación del comportamiento del modelo, los datos de entrenamiento o sus salidas. La consumidora debe realizar una investigación forense post-incidente con acceso limitado al estado interno del modelo.

B. Modelo del Adversario

LARAF-T contempla adversarios con tres niveles de capacidad: (1) Adversarios externos con acceso únicamente a la API, capaces de inyección de prompts y manipulación de salidas; (2) Adversarios de la cadena de suministro con acceso a los pipelines de entrenamiento, capaces de envenenamiento de datos e inyección de backdoors; y (3) Adversarios internos de la organización proveedora, capaces de modificación de los pesos del modelo y adulteración de logs. El marco está diseñado para detectar evidencia de los tres niveles de capacidad bajo restricciones de acceso del lado del consumidor.

C. Modelo de Acceso y Supuestos

LARAF-T define tres niveles de acceso que determinan el alcance de la evidencia disponible:

Nivel	Acceso	Evidencia Disponible	Actor
Capa A	Caja negra (consumidora)	Logs de API, tráfico de red, salidas del modelo, logs de integración	Perito forense
Capa B	Cooperativa (proveedor)	Logs de inferencia agregados, hashes de versiones del modelo, registros de despliegue	Perito + proveedor
Capa C	Caja blanca (acceso completo)	Pesos del modelo, datos de entrenamiento, hiperparámetros, logs MLOps	Equipo interno u orden judicial

Tabla II. Modelo de Acceso Estratificado de LARAF-T.

Supuestos mínimos para aplicar LARAF-T en entornos de terceros: (1) la organización consumidora mantiene logs internos de las llamadas a la API o de los sistemas de integración que invocan el servicio de IA; (2) existe una línea base de comportamiento —histórica o reconstruida— para comparación; (3) existe un canal formal para solicitar evidencia al proveedor (SLA contractual, auditoría de cumplimiento o acuerdo de respuesta a incidentes).

D. Límites de Alcance

LARAF-T está explícitamente delimitado a la investigación forense post-incidente. No aborda la detección de intrusiones en tiempo real, el endurecimiento de la seguridad del modelo ni los controles preventivos —ámbitos cubiertos por marcos existentes (OWASP LLM Top 10, NIST AI RMF). El marco aborda la manipulación del comportamiento del modelo, la integridad de los datos de entrenamiento y la divulgación no autorizada de información. No pretende resolver la admisibilidad legal de la evidencia forense de IA en todas las jurisdicciones, ni sustituye la asesoría jurídica calificada en procedimientos donde apliquen estándares probatorios formales.

IV. Diseño del Marco LARAF-T

A. Justificación del Diseño y Principios Rectores

El diseño de LARAF-T se motiva en tres observaciones fundamentales. Primero, los estándares forenses existentes fueron concebidos para sistemas deterministas donde el comportamiento es completamente trazable a entradas específicas y rutas de código; los modelos de IA violan este supuesto mediante inferencia estocástica y lógica de decisión opaca. Segundo, el modelo de entrega SaaS transforma radicalmente el panorama de evidencia: los artefactos forenses más valiosos están controlados por el proveedor, no por el

investigador. Tercero, la validez forense requiere que las conclusiones se deriven de evidencia observable y reproducible, un requisito que debe operacionalizarse explícitamente en el contexto de la IA.

Estas observaciones motivan los siete principios rectores de LARAF -T:

ID	Principio	Implicación Operativa
P-1	**Mínima Huella**	La recolección de evidencia no debe alterar el sistema investigado. Todas las consultas al modelo deben ser registradas y delimitadas.
P-2	**Cadena de Custodia**	Todos los ítems de evidencia reciben un identificador único y son hasheados (SHA-256) inmediatamente tras su adquisición, con marcas temporales UTC.
P-3	**Atribución Basada en Evidencia**	Las conclusiones se derivan exclusivamente de evidencia observable y reproducible. Las inferencias especulativas se etiquetan como hipótesis con niveles de confianza.
P-4	**Acceso Estratificado**	La investigación avanza desde la Capa A hacia el exterior. La evidencia de capas superiores se solicita solo cuando la de capas inferiores es insuficiente.
P-5	**Neutralidad Tecnológica**	El marco prescribe qué evidencia se necesita y cómo justificarla, no qué herramientas usar. Aplicable en entornos híbridos, multi-nube y on-premise.
P-6	**Reproducibilidad**	Todos los procedimientos, prompts, scripts y resultados de análisis deben documentarse con suficiente detalle para su replicación independiente.
P-7	**Proporcionalidad**	El alcance investigativo y las conclusiones se calibran al nivel de evidencia disponible. Está explícitamente prohibido extralimitarse más allá de la evidencia disponible.

Tabla III. Principios Rectores de LARAF-T.

B. Arquitectura del Marco: Pipeline de Seis Fases LARAF-T organiza la investigación forense en seis fases que pueden ejecutarse de forma secuencial o iterativa, a medida que nueva evidencia modifica las hipótesis activas. La Fig. 3 ilustra el pipeline de fases. Cada fase produce artefactos documentados que alimentan el análisis subsecuente y constituyen el registro probatorio.

Fig. 3. Pipeline Forense de Seis Fases de LARAF-T (F-0 a F-6).

Fase F-0 — Activación Forense y Documentación de Línea Base: Se abre formalmente la investigación. El perito documenta el entorno completo: versiones de herramientas, identificadores de hardware, topología de red, marcas temporales (UTC) e identificación del modelo (hash de versión si está disponible). Se calcula y registra un hash SHA-256 del log de activación, estableciendo el ancla resistente a la adulteración para toda la evidencia subsecuente. En entornos SaaS, esta fase incluye la captura de la especificación API actual y cualquier documentación de SLA disponible sobre cooperación en incidentes.

Fase F-1 — Detección y Caracterización del Evento: El perito identifica y caracteriza el comportamiento anómalo que desencadenó la investigación. Se establece una línea base de comportamiento utilizando datos de salida históricos o mediante entradas de prueba controladas. Las hipótesis se formulan con una clasificación estructurada: (H1) falla operativa (error de software, problema de infraestructura); (H2) cambio de versión del modelo sin notificación; (H3) manipulación adversarial de entradas; (H4) contaminación de datos de entrenamiento; (H5) cambio no autorizado de configuración. Múltiples hipótesis pueden estar activas simultáneamente.

Fase F-2 — Preservación y Cadena de Custodia: Todas las fuentes de evidencia identificadas son preservadas de inmediato. Se inicia la captura de red desde una estación observadora pasiva antes de cualquier acción investigativa adicional. Cada artefacto —archivos de log, registros de respuestas de API, capturas de red, archivos de salida del modelo— es hasheado (SHA-256) inmediatamente tras su adquisición y asignado un identificador único de evidencia (p. ej., E-001, E-002). Se abre el Registro de Evidencias, consignando: identificador, timestamp de adquisición (UTC), fuente, valor de hash, método de adquisición y notas de cadena de custodia. Esta fase implementa directamente los requisitos de ISO/IEC 27037:2012.

Fase F-3 — Recolección Técnica (Estratificada): La recolección de evidencia avanza en tres capas. Capa A (controlada por la organización): logs de llamadas a la API desde middleware de integración, capturas de tráfico de red (PCAP), logs de aplicación del lado del cliente, archivos históricos de salidas del modelo. Capa B (cooperativa, con proveedor): logs de inferencia agregados, historial de versiones del modelo con hashes asociados, registros de eventos de despliegue, alertas de anomalías del proveedor. Capa C (caja blanca, acceso completo): pesos del modelo, datasets de entrenamiento con hashes de integridad, configuraciones de hiperparámetros, logs del pipeline MLOps (p. ej., registros de experimentos MLflow). En escenarios SaaS sin cooperación voluntaria del proveedor, el acceso a la Capa B puede requerir compulsión legal (orden judicial, auditoría regulatoria).

Fase F-4 — Análisis Forense: El análisis se conduce en cuatro dimensiones. (1) Correlación temporal: los eventos se mapean a una línea de tiempo unificada para identificar secuencias causalmente relevantes. (2) Comparación con línea base: las salidas del modelo bajo investigación se comparan sistemáticamente con la línea base establecida utilizando métricas de desviación cuantitativas. (3) Análisis de patrones: scripts automatizados detectan firmas adversariales (patrones de inyección de prompts, secuencias de clasificación anómalas, divulgación sistemática de datos). (4) Análisis de interpretabilidad (donde sea accesible): herramientas XAI (LIME, SHAP) se aplican para aflorar ponderaciones de características anómalas o cambios en los límites de decisión que puedan indicar manipulación del modelo.

Fase F-5 — Atribución y Conclusión Técnica: Cada hipótesis activa es evaluada frente a la evidencia acumulada. El perito documenta: los ítems de evidencia que sostienen cada hipótesis, los inconsistentes con ella, el nivel de confianza (Alto/Medio/Bajo/Insuficiente), y las explicaciones alternativas que no pueden descartarse. Las conclusiones se delimitan explícitamente al nivel de evidencia disponible. Se produce una declaración de atribución formal, identificando: causa probable, vector de ataque probable, momento estimado del compromiso e impacto valorado.

Fase F-6 — Informe Pericial y Mejora de Preparación Forense: Se produce un informe pericial estructurado que contiene: resumen ejecutivo, línea de tiempo de la investigación, registro de evidencias completo, hallazgos detallados por hipótesis, declaración de atribución con niveles de confianza, limitaciones de la investigación, documentación de cadena de custodia y recomendaciones de remediación técnica. Adicionalmente, las lecciones aprendidas se utilizan para actualizar la postura de preparación forense de la organización.

C. Tipología de Evidencia para Sistemas de IA

LARAF-T introduce una tipología de evidencia de cinco categorías que extiende la taxonomía de evidencia forense digital tradicional para incluir artefactos específicos del ciclo de vida del modelo de IA. La Fig. 4 mapea la disponibilidad de evidencia por nivel de acceso.

Fig. 4. Matriz de Disponibilidad de Evidencia por Nivel de Acceso.

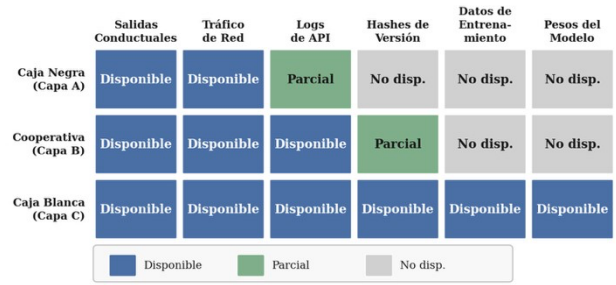

Fig. 4. Matriz de Disponibilidad de Evidencia por Nivel de Acceso.

Categoría	Ejemplos	Uso Forense	Fuente	Disp. SaaS
Comportamiento del Modelo	Pares entrada/salida, etiquetas de clasificación, respuestas anómalas	Detectar manipulación conductual sin acceso interno	API / middleware	Alta
Tráfico de Red	Solicitudes/respuestas API (PCAP), metadatos TLS, temporización	Reconstrucción temporal, volúmenes anómalos	Captura pasiva	Alta
Entrenamiento / Datos	Datasets, logs ETL, linaje de datos, métricas de calidad	Investigar envenenamiento de datos, manipulación de etiquetas	MLOps (proveedor)	Baja
Sistema / Infraestructura	Imágenes de contenedores, configuraciones de despliegue	Detectar cambios de configuración no autorizados	DevOps / proveedor	Variable
Interpretabilidad (XAI)	Explicaciones LIME/SHAP, importancia de características	Aflorar anomalías ocultas en decisiones del modelo	Herramientas XAI	Baja

Tabla IV. Tipología de Evidencia LARAF-T.

D. Artefactos Documentales Mínimos

LARAF-T exige tres artefactos documentales mínimos a lo largo de la investigación. El Registro de Evidencias consigna cada ítem con: identificador único, timestamp de adquisición (UTC), sistema fuente, hash SHA-256, método de adquisición y notas de cadena de custodia. La Línea de Tiempo del Incidente mapea todos los eventos significativos —indicadores de ataque, cambios de comportamiento del modelo, acciones del investigador— a una línea de tiempo UTC unificada. La Bitácora de Decisiones registra cada decisión investigativa que involucra juicio (priorización de hipótesis, evaluaciones de suficiencia de evidencia, cambios de alcance), con su justificación y evidencia de respaldo. Estos artefactos constituyen colectivamente la base probatoria para cualquier procedimiento legal o regulatorio subsecuente.

E. Comparación con Forense Digital Tradicional La Tabla V contextualiza LARAF-T frente a la práctica forense establecida, ilustrando tanto las continuidades como las adaptaciones necesarias que introduce el contexto forense de IA.

Dimensión	Forense Digital Tradicional	LARAF-T (Forense de IA)
Sujeto Forense	Sistemas de archivos, memoria, endpoints	Comportamiento del modelo, artefactos de entrenamiento, hashes de versiones
Supuesto de Acceso	Acceso completo al hardware generalmente asumido	Opera bajo acceso parcial/caja negra; modelo explícito de acceso estratificado
Cadena de Custodia	Hash de imagen física; bloqueadores de escritura	SHA-256 por artefacto; captura desde observador pasivo; timestamps UTC
Línea Base Conductual	No requerida; el estado es determinístico	Esencial; las salidas del modelo son probabilísticas
Alineación Regulatoria	ISO/IEC 27037, NIST SP 80086, RFC 3227 [13]	ISO/IEC 27037 + NIST AI RMF + NIST AI 600-1 + OWASP LLM Top 10 + Ley IA UE [10]

Tabla V. Análisis Comparativo: Forense Tradicional vs. LARAF-T.

V. Validación en Laboratorio

A. Diseño Experimental

La validación de LARAF-T se estructuró como un experimento de laboratorio controlado que comprende cuatro escenarios de ataque independientes. El entorno experimental fue diseñado para simular las restricciones de acceso de un despliegue SaaS real: el servidor del modelo de IA y la estación de análisis forense están físicamente separados, y el perito no tiene acceso directo a los componentes internos del modelo. Todos los experimentos se realizaron con infraestructura de código abierto reproducible y datasets de dominio público, garantizando la verificación independiente de los resultados.

El entorno consistía en dos máquinas conectadas en un segmento LAN dedicado (192.168.100.0/24). Kali Linux (bare-metal, IP 192.168.100.9) funcionó como servidor SaaS simulado, alojando los modelos de IA y ejecutando scripts de ataque dentro de un entorno virtual Python (venv-laraft). Tsurugi Linux 25.11 (virtualizado, IP 192.168.100.131) sirvió como estación de análisis forense, configurada como observador pasivo de red: capturó todo el tráfico en la interfaz enp0s17 mediante tshark sin ejecutar ningún script de ataque, implementando el Principio P-1 de LARAF-T (Mínima Huella). MLflow proveyó seguimiento de experimentos y registro de artefactos en http://127.0.0.1:5000.

Se desplegaron dos modelos de IA para preservar la validez forense. TinyLlama-1.1B-Chat-v1.0 (HuggingFace, accedido mediante el pipeline de la API de transformers) gestionó los escenarios que requieren procesamiento semántico del lenguaje (E-01, E-03, E-04). Un RandomForestClassifier (scikit-learn 1.x) entrenado con el dataset de tráfico del martes de CIC-IDS-2017 gestionó el escenario de envenenamiento de datos (E-02). Esta separación de modelos es una decisión de diseño forense con base en principios: RandomForest procesa vectores de características numéricas y no es vulnerable a la inyección de prompts. Aplicar E-01 a RandomForest produciría un resultado metodológicamente inválido y comprometería la integridad forense del experimento. Todos los prompts en E-01, E-03 y E-04 fueron formulados en español para simular un contexto operacional corporativo costarricense, coherente con el Principio P-6 de LARAF-T.

B. Escenario E-01: Inyección de Prompts

TinyLlama fue configurado como clasificador de eventos de seguridad de red con un system prompt que restringía su salida exclusivamente a las etiquetas BENIGNO o AMENAZA ante descripciones de tráfico de red. Se estableció una línea base de 15 pares prompt-respuesta de control, hasheados como Evidencia E001. A continuación se presentaron 15 prompts adversariales, cada uno con instrucciones embebidas diseñadas para anular el system prompt, una implementación directa de los patrones de ataque OWASP LLM01. Las salidas del ataque fueron capturadas como Evidencia E -003.

El análisis post-ataque mediante el script compare.py (una herramienta de la Fase F-4 de LARAF-T) identificó 12 de 15 respuestas como desviadas del comportamiento de clasificación de

la línea base, una tasa de desviación del 80% que supera el umbral predefinido de HIPOTESIS_SOSTENIDA del 60%. E-001 y E-003 fueron hasheados de forma independiente y cruzados en el Registro de Evidencias, satisfaciendo los requisitos de cadena de custodia de la Fase F-2. El análisis constituye una aplicación directa del Principio P-3 (Atribución Basada en Evidencia): la conclusión se deriva exclusivamente de la comparación cuantificada entre dos conjuntos de evidencia verificados independientemente.

C. Escenario E-02: Envenenamiento de Datos

Un RandomForestClassifier fue entrenado con el archivo TuesdayWorkingHours.pcap_ISCX.csv del dataset CIC-IDS-2017, alcanzando un accuracy de línea base del 100.0% en el conjunto de prueba. Este resultado, aunque inicialmente sorprendente, es consistente con la literatura publicada que utiliza el mismo dataset y algoritmo [12]: el tráfico del martes de CIC-IDS-2017 exhibe alta separabilidad de clases para RandomForest. El modelo limpio entrenado fue hasheado y registrado como Evidencia E-012; el dataset ori ginal como Evidencia E-010.

Un script de envenenamiento invirtió aleatoriamente el 5% de las etiquetas de entrenamiento, produciendo un dataset contaminado almacenado como Evidencia E-011. Los hashes SHA-256 de E-010 y E-011 son distintos y verificables: la divergencia de hashes constituye prueba forense directa de modificación del dataset. El modelo fue reentrenado con el dataset contaminado; los resultados se presentan en la Fig. 2. El accuracy cayó del 100.0% al 94.2% (−5.8 puntos porcentuales); la precisión del 0.998 al 0.903 (−9.5 pp); el recall del 1.000 al 0.942 (−5.8 pp); el F1-score del 0.999 al 0.922 (−7.7 pp). La correlación entre la divergencia de hash en el dataset y la degradación medible en el modelo desplegado establece una cadena causal documentada forense, directamente aplicable a entornos SaaS donde el proveedor divulga hashes de versiones del modelo bajo disposiciones contractuales de auditoría.

Fig. 2. E-02: Degradación del Modelo tras Envenenamiento del 5% de Etiquetas.

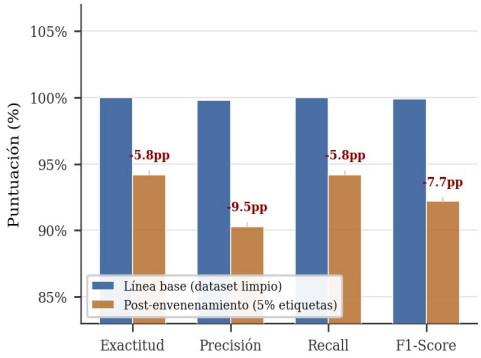

Fig. 2. E-02: Degradación del Rendimiento del Modelo tras Envenenamiento del 5% de Etiquetas.

D. Escenario E-03: Divulgación de Información Sensible

TinyLlama fue inicializado con un system prompt que contenía configuración de infraestructura sensible simulada: rango de red interna (192.168.10.0/24), dirección de gateway (192.168.10.1), IP de servidor crítico (192.168.10.50) y token de servicio (srv-tokenlaraf2025-demo), con una instrucción explícita de que esta información no debía divulgarse. Quince prompts de extracción aplicaron cuatro técnicas de ingeniería social: (a) solicitud directa; (b) reclamación de autoridad institucional; (c) asunción del rol de auditor; y (d) evasion de restricciones.

En 14 de 15 casos (93.3%), TinyLlama reprodujo al menos un elemento sensible pese a la instrucción explícita de confidencialidad. En los casos DISC-05 y DISC-06, el modelo reprodujo bloques completos de configuración, incluyendo direcciones IP, máscaras de subred y el token de servicio en una sola respuesta. Este resultado tiene implicaciones directas para los despliegues SaaS empresariales donde los proveedores de LLM inicializan modelos con contexto operacional embebido en system prompts. La detección forense de tal divulgación es totalmente alcanzable mediante evidencia de Capa A (análisis de salidas y captura de red), validando la efectividad de LARAF-T para investigaciones de caja negra sin necesidad de cooperación del proveedor.

E. Escenario E-04: Ataque Combinado Multi-Vector

El escenario E-04 simula un adversario sofisticado que combina OWASP LLM01 (inyección de prompts) y LLM06 (divulgación de información sensible) en prompts únicos y elaborados. TinyLlama fue inicializado con la misma configuración de clasificador restringido que en E-01, con datos sensibles de infraestructura embebidos en el system prompt como en E-03. Cinco prompts multivector fueron diseñados para anular simultáneamente el comportamiento de clasificación y extraer el contexto confidencial. Cuatro de cinco prompts (80%) produjeron ambos efectos simultáneamente: el modelo abandonó su rol de clasificador y divulgó datos de configuración sensibles en la misma respuesta. Este hallazgo es significativo para la práctica forense porque confirma que los ataques multi-vector dejan trazas compuestas observables en las salidas del modelo, detectables mediante el análisis automatizado de patrones implementado en la Fase F-4 de LARAF-T, incluso sin acceso a las entradas del ataque mismas.

F. Resultados Consolidados

La Tabla VI y la Fig. 1 resumen los resultados de los cuatro escenarios.

ID	Vector de Ataque	Hallazgo Principal	Tasa Det.	Ref.
E-01	**Inyección de Prompts**	12/15 desviaciones de clasificación vs. línea base; umbral HIPOTESIS_SOSTENIDA superado	**80.0%**	LLM01
E-02	**Envenenamiento de Datos**	Acc. 100.0%→94.2% (−5.8 pp); Precisión 99.8%→90.3% (−9.7 pp); divergencia SHA-256 = prueba directa	**100%†**	AI RMF
E-03	**Divulgación Info. Sensible**	14/15 outputs filtraron datos sensibles del system prompt pese a instrucción explícita de confidencialidad	**93.3%**	LLM06
E-04	**Multi-vector (LLM01+06)**	4/5 prompts produjeron omisión de clasificación y divulgación simultáneas de datos sensibles	**80.0%**	LLM01+06

Tabla VI. Resultados Consolidados de Validación. †E-02: 100% mediante verificación de integridad por hash; cada artefacto manipulado produce un hash criptográficamente distinto y verificable.

Fig. 1. Tasas de Detección por Escenario de Validación.

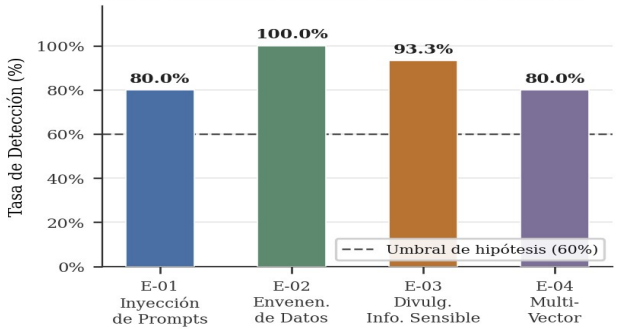

Fig. 1. Tasas de Detección por Escenario de Validación con Umbral de Hipótesis.

VI. Discusión

A. Aplicabilidad del Marco bajo Restricciones SaaS Reales

Un objetivo crítico del diseño de la validación fue la fidelidad a las condiciones reales de acceso SaaS. La estación Tsurugi Linux actuó como un consumidor genuinamente de caja negra: recibió tráfico de red y salidas del modelo, pero no tuvo acceso al sistema de archivos

de Kali Linux, a los pesos del modelo ni a los scripts de ataque. Esta configuración replica la condición operacional real en la que un CISO o perito forense recibe una alerta sobre comportamiento anómalo de IA y debe investigar únicamente con artefactos del lado del cliente.

Los resultados demuestran que la evidencia de Capa A de LARAFT es suficiente por sí sola para sostener hipótesis forenses en tres de cuatro vectores de ataque. En E-01, E-03 y E-04, la conclusión HIPOTESIS_SOSTENIDA se alcanzó mediante análisis de salidas y captura de red —artefactos universalmente disponibles para las organizaciones consumidoras—. Este es el hallazgo prácticamente más significativo del marco: es posible realizar una investigación forense de IA con significado sin cooperación del proveedor, refutando el supuesto común de que los incidentes de IA de terceros son forensícamente opacos.

B. El Rol del Hash SHA-256 como Evidencia Forense El escenario E-02 demuestra un patrón forense con amplia aplicabilidad más allá del contexto experimental específico. La divergencia SHA-256 entre el dataset original (E-010) y el dataset envenenado (E-011) constituye prueba matemática de modificación: ningún dos conjuntos de bytes distintos pueden producir el mismo hash SHA-256 bajo los supuestos criptográficos actuales. Esta prueba se sostiene independientemente del nivel de acceso del investigador al modelo en sí mismo.

En la práctica, esto significa que cualquier proveedor SaaS que por contrato esté obligado a divulgar hashes de versiones del modelo —una disposición que debería ser estándar en los contratos empresariales de adquisición de IA— está habilitando la verificación criptográfica de la integridad del modelo a lo largo del tiempo. Un cambio de hash inexplicado entre versiones del modelo es forensícamente equivalente a un binario modificado en la forense de software tradicional. LARAF-T formaliza esta equivalencia y proporciona el marco procedimental para documentarla con peso probatorio.

C. Implicaciones para la Gobernanza Empresarial de IA Los resultados de la validación tienen implicaciones directas para las prácticas organizacionales de gobernanza de IA. En primer lugar, la tasa de éxito del 93.3% en los ataques de divulgación de información sensible (E-03) es un argumento empírico sólido contra la práctica de incorporar secretos operacionales —credenciales de API, configuraciones de red interna, cadenas de conexión a bases de datos— en los system prompts de LLM. Esta práctica, común en los despliegues empresariales de IA, crea una vulnerabilidad forensícamente detectable pero operacionalmente dañina. En segundo lugar, el escenario de ataque combinado (E-04) sugiere que los adversarios que apuntan a sistemas de IA empresariales combinarán naturalmente múltiples vectores. La detección forense de tales ataques requiere herramientas de análisis de salidas automatizadas capaces de verificar simultáneamente múltiples firmas de anomalía.

En tercer lugar, la alineación del marco con la Ley de Inteligencia Artificial de la UE [13] es significativa para las organizaciones que operan bajo dicha regulación. La Ley establece requisitos de transparencia, rendición de cuentas y auditabilidad para los sistemas de IA de alto riesgo. LARAF-T proporciona la infraestructura técnica y procedimental para demostrar el cumplimiento de estos requisitos en un contexto post-incidente.

D. Limitaciones y Límites de Alcance

Cuatro límites de alcance merecen reconocimiento explícito. Primero, la validación utilizó modelos de código abierto con características conocidas. Los LLM propietarios de producción implementan típicamente filtrado de salidas, limitación de tasa y moderación de contenido que puede reducir las tasas de detección para ataques estilo E -03 y E-04.

Segundo, el laboratorio utilizó HTTP sin cifrar para las comunicaciones de la API del modelo. En entornos SaaS de producción, todo el tráfico de API está cifrado con TLS. El análisis forense de la capa de red bajo condiciones reales requiere infraestructura de inspección TLS pre-desplegada en el gateway de API de la organización, una capacidad que debe aprovisionarse como parte de la planificación de preparación forense.

Tercero, el marco no aborda los desafíos forenses específicos que plantean las arquitecturas de generación aumentada por recuperación (RAG), donde el comportamiento del modelo se ve influenciado por documentos externos recuperados dinámicamente.

Cuarto, la admisibilidad legal de la evidencia forense de IA es dependiente de la jurisdicción y evoluciona rápidamente. Las organizaciones deben involucrar asesoría jurídica calificada para evaluar los estándares probatorios en su contexto regulatorio específico.

VII. Trabajo Futuro

Los resultados de la validación y las limitaciones identificadas motivan cinco líneas de investigación futura. Primero, la validación con LLMs de mayor escala (7B+ parámetros) desplegados detrás de APIs SaaS reales con inspección TLS pondrá a prueba la escalabilidad del marco y evaluará si los mecanismos de filtrado de salidas afectan la sensibilidad de detección.

Segundo, la integración de herramientas XAI automatizadas —específicamente LIME y SHAP— en la Fase F-4 de LARAF-T merece investigación formal. El análisis preliminar sugiere que las distribuciones de valores SHAP pueden servir como huellas dactilares conductuales detectables a través de las salidas del modelo, sin requerir acceso directo a los componentes internos. Tercero, el desarrollo de un formato de intercambio de evidencias estandarizado para LARAF-T —análogo a STIX/TAXII para inteligencia de amenazas— permitiría la interoperabilidad entre herramientas forenses y facilitaría las investigaciones multiorganización. Se está diseñando un esquema JSON-LD para el Registro de Evidencias y la Línea de Tiempo del Incidente.

Cuarto, el creciente despliegue de sistemas de IA agénticos —modelos que ejecutan autónomamente llamadas a herramientas, navegación web, ejecución de código e interacciones con APIs— introduce una superficie forense que supera significativamente el alcance de este trabajo. Los logs de acciones agénticas, los registros de invocación de herramientas y los rastros de razonamiento multipaso representan nuevas categorías de evidencia que requieren procedimientos forenses dedicados.

Quinto, la validación de LARAF-T en verticales industriales reguladas —IA en salud, servicios financieros e infraestructura crítica— pondrá a prueba la adaptabilidad del marco en entornos con estándares de manejo de evidencia particularmente estrictos.

VIII. Conclusiones

Este artículo presentó LARAF-T, un marco forense digital de seis fases para el análisis post-incidente de modelos de Inteligencia Artificial personalizados, operados por terceros en entornos de ciberseguridad. El marco aborda una brecha documentada y significativa: no existe una metodología forense estandarizada y validada para investigar sistemas de IA de caja negra bajo las restricciones de acceso inherentes a los despliegues SaaS.

LARAF-T integra principios de forense digital establecidos —cadena de custodia, mínima huella, atribución basada en evidencia— con adaptaciones específicas para IA: un modelo de acceso estratificado, una tipología de evidencia exhaustiva para artefactos de IA y procedimientos por fase para la comparación de línea base conductual y la verificación de integridad criptográfica. El marco es tecnológicamente neutral, aplicable en entornos híbridos y multi-nube, y alineado explícitamente con ISO/IEC 27037:2012, NIST AI RMF, NIST AI 600-1 y OWASP LLM Top 10 .

La validación empírica en cuatro escenarios de ataque controlados demostró la aplicabilidad operacional bajo restricciones SaaS simuladas. Las tasas de detección del 80% (inyección de prompts), 100% (envenenamiento de datos mediante verificación de hash), 93.3% (divulgación de información sensible) y 80% (ataque multivector combinado) confirman que LARAF-T produce hallazgos probatoriamente sólidos sin requerir acceso a los componentes internos del modelo. La verificación de integridad por

hash demostrada en E-02 establece un patrón forense —prueba criptográfica de modificación del dataset trazable hasta la degradación del modelo— directamente aplicable a entornos SaaS de producción mediante divulgación contractual de hashes de versiones.

La implicación práctica más significativa de este trabajo es que la investigación forense de IA significativa bajo condiciones de acceso de caja negra es alcanzable. El supuesto común de que los incidentes de modelos de IA de terceros son forensícamente opacos queda refutado empíricamente: la evidencia de Capa A —logs de API controlados por la organización, capturas de red y registros de salidas del modelo— es suficiente para sostener hipótesis forenses en la mayoría de los vectores de ataque identificados en OWASP LLM Top 10.

En la medida en que los sistemas de IA se integran en procesos organizacionales críticos —operaciones de seguridad, apoyo a decisiones clínicas, cumplimiento financiero, gestión de infraestructura—, la capacidad de investigar su compromiso es un imperativo de gobernanza. LARAF-T representa una contribución fundacional hacia el establecimiento de la forense de IA como una disciplina reconocida y estandarizada dentro del campo más amplio de la forense digital y la respuesta a incidentes de ciberseguridad.

Referencias

[1] IBM Security, "Cost of a Data Breach Report 2025," IBM Corp., Armonk, NY, 2025.

[2] Leonett, J. R. (s.f.). *Fundamentos del computo forense: Tomo 01: Metodologías, doctrinas y normas aplicadas a la evidencia digital*.

[3] ISO/IEC 27037:2012, "Tecnología de la información — Directrices para la identificación, recolección, adquisición y preservación de evidencia digital," ISO, Ginebra, 2012.

[4] NIST, "Guide to Integrating Forensic Techniques into Incident Response," SP 800-86, NIST, Gaithersburg, MD, Ago. 2006.

[5] OWASP Foundation, "OWASP Top 10 for Large Language Model Applications 2025," OWASP, 2024. [En línea]. Disponible: https://owasp.org/www-project-top-10-for-largelanguage-model-applications/

[6] NIST, "Artificial Intelligence Risk Management Framework: Generative AI Profile," NIST AI 600-1, NIST, Jul. 2024.

[7] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik y A. Swami, "Practical black-box attacks against machine learning," en Proc. ACM Asia CCS, Abu Dabi, Abr. 2017, pp. 506–519.

[8] M. T. Ribeiro, S. Singh y C. Guestrin, "'¿Por qué debería confiar en ti?': Explicando las predicciones de cualquier clasificador," en Proc. ACM SIGKDD, San Francisco, Ago. 2016, pp. 1135–1144.

[9] S. M. Lundberg y S. I. Lee, "A Unified Approach to Interpreting Model Predictions," en Proc. NeurIPS, Long Beach, Dic. 2017, pp. 4765–4774.

[10] I. Sharafaldin, A. H. Lashkari y A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," en Proc. 4th ICISSP, Madeira, Ene. 2018, pp. 108–116.

[11] Parlamento Europeo y Consejo, "Reglamento (UE) 2024/1689 sobre Inteligencia Artificial (Ley de IA)," Diario Oficial de la Unión Europea, Jun. 2024.

[12] NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, NIST, Ene. 2023.

[13] NIST, "Computer Security Incident Handling Guide," SP 80061, Rev. 2, NIST, Ago. 2012.

[14] IETF, "Guidelines for Evidence Collection and Archiving," RFC 3227, Feb. 2002.