https://doi.org/10.69639/arandu.v13i3.2533

Análisis del uso de herramientas basadas en Procesamiento del lenguaje natural (NLP) y Grandes modelos de lenguaje (LLM) en la reducción de complejidad técnica de ciberataques ejecutados por actores con conocimientos limitados

Analysis of the Use of Natural Language Processing (NLP) and Large Language Models (LLM) -Based Tools in Reducing the Technical Complexity of Cyberattacks Executed by Actors with Limited Technical Knowledge

Cristhian Oswaldo Vera Segarra
e1720252574@uisrael.edu.ec
https://orcid.org/0009-0008-0813-7550
Universidad Tecnológica Israel
Quito – Ecuador

Juan Joel Caillagua Taco
e1724168784@uisrael.edu.ec
https://orcid.org/0009-0005-3307-1563
Universidad Tecnológica Israel
Quito – Ecuador

Elian Ricardo Farinango Cevallos
e1005253453@uisrael.edu.ec
https://orcid.org/0009-0003-5449-5854
Universidad Tecnológica Israel
Quito – Ecuador

Artículo recibido: 10 julio 2026- Aceptado para publicación:16 agosto 2026
Conflictos de intereses: Ninguno que declarar.

RESUMEN

El avance de las tecnologías basadas en inteligencia artificial y de los modelos de Natural Language Processing (NLP) y Large Language Models (LLM), han generado oportunidades y desafíos en el ámbito de la ciberseguridad, sobre todo en su utilización con fines ofensivos. El objetivo de esta investigación es analizar como los modelos NLP y LLM están siendo usados como herramientas en actividades ofensivas y determinar su impacto en la reducción de la complejidad técnica para le ejecución de ciberataques. Para el estudio se desarrolló una revisión sistemática de literatura mediante la recopilación y análisis de artículos científicos publicados entre 2023 y 2026. Los resultados evidenciaron el uso de modelos GPT, LLaMA, Gemini, Claude y DeepSeek en actividades ofensiva, se identificó que estas tecnologías reducen significativamente la barrear técnica para ejecutar determinadas actividades ofensivas, incrementado la automatización, personalización y escalabilidad. El estudio llega a tener conclusiones muy acetadas sobre el trasfondo de uso de NLP y LLM y como estas están

transformando el panorama de la ciberseguridad ofensiva facilitando las capacidades avanzadas de ataque para el uso de actores con conocimientos limitados.

Palabras claves: inteligencia artificial, procesamiento de lenguaje natural, modelos de lenguaje de gran escala, ciberataques, phishing

ABSTRACT

The advancement of technologies based on Artificial Intelligence (AI), Natural Language Processing (NLP), and Large Language Models (LLMs) has created both opportunities and challenges in the field of cybersecurity, particularly regarding their use for offensive purposes. The objective of this research is to analyze how NLP and LLM models are being used as tools in offensive cybersecurity activities and to determine their impact on reducing the technical complexity required to execute cyberattacks. To achieve this objective, a systematic literature review was conducted through the collection and analysis of scientific articles published between 2023 and 2026. The results revealed the use of models such as GPT, LLaMA, Gemini, Claude, and DeepSeek in offensive activities. Furthermore, the findings indicate that these technologies significantly reduce the technical barriers associated with conducting certain offensive operations, while increasing the automation, personalization, and scalability of cyberattacks. The study provides valuable insights into the underlying use of NLP and LLM technologies and demonstrates how they are transforming the landscape of offensive cybersecurity by enabling advanced attack capabilities to be leveraged by actors with limited technical knowledge.

Keywords: artificial intelligence, natural language processing, large language models, cyberattacks, phishing

INTRODUCCIÓN

La transformación digital global y el crecimiento acelerado de tecnologías basadas en inteligencia artificial han generado cambios significativos dentro de diferentes áreas de la informática, economía, desarrollo de software, en el ámbito de la ciberseguridad y la seguridad informática.

Entre las tecnologías actuales basadas en IA se encuentran las herramientas como el Natural Language Processing y Large Language Models, las cuales permiten interactuar mediante lenguaje natural para realizar tareas relacionadas con generación de código, automatización de procesos, análisis de información y asistencia técnica.

La accesibilidad de estas tecnologías ha permitido que usuarios con conocimientos técnicos limitados puedan utilizar herramientas avanzadas de inteligencia artificial para ejecutar actividades que anteriormente requerían mayores niveles de experiencia técnica. Si bien un Machine Learning (ML) puede garantizar y mejorar la seguridad en la red, también puede ser usado en su contra como un arman, los ataques basados en ML son menos predecibles y con los promts correctos se puede genera código que puede ser usado para atacar un sistema. (Czaja et al., 2025)

La proyección de la inseguridad cibernética debido al aumento de cibercrímenes con el uso de tecnologías disruptivas genera una preocupación justificada en la inminente evolución de los ciberataques con el uso de estas tecnologías disruptivas. Los debates sobre estos riesgos han generado discusiones desde diferentes perspectivas, entendiendo que es un peligro tanto en la actualidad como en el futuro, y como sociedad debemos abordar estos temas. (Jiménez, 2025)

Justamente como hace alusión Jones en sus conclusiones de investigación, desde una perspectiva de seguridad informática ofensiva el creciente uso de LLM en la ciberseguridad plantea nuevas incógnitas, y están apareciendo nuevas amenazas que pueden ser aprovechadas por actores maliciosos. (Jones et al., 2025)

Este estudio busca identificar herramientas LLM y NLP utilizadas de manera ofensiva para ejecutar automatizaciones, análisis y asistencia técnica en la reducción de la complejidad requerida para realizar actividades ofensivas, que puedan ser usadas por actores maliciosos con conocimientos limitados.

Los LLMs están posibilitando e incrementado las amenazas cibernéticas, como lo muestra la revisión realizada por (Alqahtani & Kumar, 2026), quien destaca que los ataques realizado con contenido generado por IA representa una diferenciación de cualidades, más que ser un impulsor incremental de técnicas que ya se conocía. En ataques conocidos como pishing, ingeniería social, en el desarrollo de malware y exploits los LLMs introducen una escalabilidad, una inteligencia contextual y adaptabilidad sin precedentes, reduciendo la barra de conocimiento técnica para los atacantes, mientras que los mecanismos de protección tradicionales pierden efectividad.

En consecuencia ¿Los LLMs está reduciendo la complejidad de ejecutar ciberataques para personas con poco conocimiento técnico?

¿Qué herramientas basadas en Natural Language Processing (NLP) y Large Language Model (LLM) o en funciones de Machine Learning están facilitando ataques a actores con conocimientos limitados? ¿Hasta qué punto la efectividad de uso de LLM y NLP en estas herramientas puede ser aprovechada para fines maliciosos?

Esta investigación permite dar un enfoque de concientización sobre la accesibilidad y usabilidad de estas tecnologías disruptivas por personas con intenciones de realizar acciones ofensivas consideradas como ciberataques, con conocimientos y experticia limitada.

METODOLOGÍA

Para el desarrollo de la investigación esta se enmarca en una revisión cualitativa de manera descriptiva, confirmada por una revisión sistemática de literaturas.

Esta revisión investigativa se desarrolla en tres etapas que permitirán revisar las innovaciones con tecnologías NLP y LLM en el marco de la ciberseguridad y las tecnologías de este tipo están dentro del ámbito de la seguridad informática ofensiva, permitiendo realizar una revisión técnica, critica y sistemática sobre las fuentes empleadas.

Fase 1: Recolección y selección de información, en esta fase se desarrollará la identificación y compilación de artículos científicos, trabajos académicos que puedan obtenerse de base de datos de ScienceDirect, ACM Digital Library y Springer Nature. El criterio de vigencia se basará en información publicada a partir del 2023 enfocada aportar contexto o información directa al uso de NLP y LLM en técnicas de ciberataques.

Fase 2: Clasificación y el análisis del material se organizar en matrices las cuales se distribuyen por metodología usada en cada estudio, principales hallazgos, y aportes alcanzados en el enfoque de estas herramientas en el uso de ataque ofensivos, después se interpretará la información de forma comparativa entre cada uno de los temas y se generar un resumen evidenciando las tendencias y puntos importantes en el contexto de la temática.

Fase 3: Síntesis, en esta última fase se llevará a cabo una síntesis holística, en la cual se evidenciará los resultados que se vinculan al contexto y a la intención de concientización, permitiendo destacar los puntos más relevantes del uso de NLP y LLM como herramientas o como parte de herramientas de ciberataques ofensivas.

Se trabajará mediante la metodología PRISMA para garantizar transparencia y el rigor de la revisión semántica.Esta metodología permitió estructura los registros, e ir segmentándolos de manera granular para que cumplieran con los criterios de inclusión en lo que respecta su relevancia, contemporaneidad y calidad respecto al detalle de las herramientas de ciberataques ofensivas.

RESULTADOS

Fase 1: Recolección y selección de información

En esta fase se realizó una recolección sistemática sobre artículos científicos y conference papers que tuvieran relación con la temática abordada, centrándose en herramientas NLP y LLM en el contexto de la ciberseguridad. A través de esta fase exploratoria y de selección, se encontraron 41 artículos publicados en bases como ScienceDirect, ACM Digital Library y Springer Nature.

En la fase de selección, se eliminaron artículos con temática duplicada que no aportasen un valor diferencial o que no contribuyeran a la idea del uso ofensivo de los LLM o Machine Learning, resaltando puntos sobresalientes. Se eliminaron 25 estudios que no cumplieron con los criterios de inclusión.

Como conclusión de esta fase, se seleccionaron 16 estudios que cumplieron los criterios de inclusión según la Tabla 1 y no entraron en los artículos excluidos según las consideraciones de la Tabla 2, asegurando una calidad científica de aportación al estudio del tema central y garantizando que la revisión sistemática cumpla con la metodología PRISMA.

Tabla 1

Pautas de inclusión aplicados en la revisión sistemática

Categoría	Descripción
Periodo de publicación	Se recolectaron únicamente los artículos publicados desde enero 2023 a mayo 2026, con el fin de garantizar una relevancia científica de estudios actualizada.
Tipo de documentos	Se revisaron artículos científicos y conference papers, seleccionando aquello que evidenciaban contribuciones significativas y alineadas con los objetivos de la investigación.
Accesibilidad	Los documentos deben ser en texto completo o contar con acceso abierto que posibilitaran su revisión, se consideraron como prioridad las publicaciones en inglés o español de las bases ScienceDirect, ACM Digital Library y Springer Nature.
Relevancia Académica	Se seleccionaron estudios relacionados con herramientas y aplicaciones ofensivas que se basaran en NLP y LLM.
Rigor Metodológico	Los artículos escogidos debían mantener un claro diseño metodológico e investigativo según el propósito de sus objetivos, aportando evidencia conceptual significativa para el tema

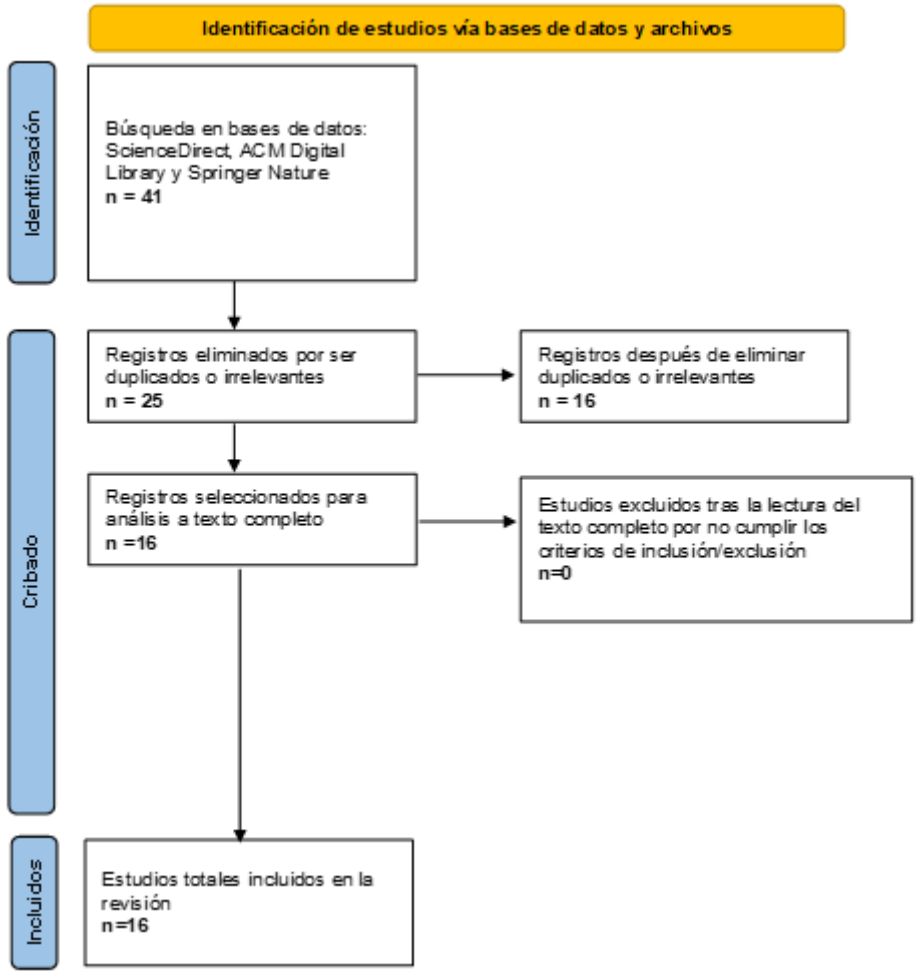
Tabla 2

Pautas de exclusión aplicado en la revisión sistemática

Categoría	Descripción
Periodos excluidos	Se excluyeron artículos anteriores a 2023, debido a que el estudio plantea ser lo más actualizado posible.

Revisión por pares	No se consideraron dentro del estudio artículos sin revisión académica o con información su validación científica.
Inaccesibilidad	Los documentos que no se presentaran con accesibilidad para su estudio fueron descartados de la revisión.
Irrelevancia Académica	Se omitieron estudios que no presentaran evidencia de uso de herramientas de ciberseguridad de forma ofensiva o que no incluyeron el contexto de ciberataque ofensivo.
Duplicidad	No se seleccionaron artículos duplicados o versiones redundantes que no dieran un enfoque distinto a las herramientas estudiadas.

Figura 1
Metodología PRISMA



Fase 2: La clasificación y el análisis del material

La revisión sistemática resume los hallazgos encontrados que respaldan el estado del arte sobre las herramientas de ciberataque ofensivo que hacen uso de LLM basados en NLP, permitiendo identificar la facilidad de uso en ciertos aspectos, considerando la poca necesidad de conocimientos técnicos avanzados.

A través del análisis de los 16 artículos científicos dentro del período revisado 2023 – 2026, se observan diversos desarrollos de técnicas de ataque ofensivo y la evolución de las mismas mediante el uso de LLM y Machine Learning. Estos nuevos desarrollos y formas de ataque permiten evidenciar un panorama variado donde los LLM aportan, como tecnología disruptiva, una mayor eficacia dentro del entorno ofensivo.

La metodología PRISMA permitió mantener un enfoque granular y objetivo en la selección de las fuentes de estudio, considerando su aporte empírico y científica sobre la temática abordada. Esto facilitó obtener conclusiones alineadas al objetivo de mantener una concientización y alerta sobre la tendencia de estas tecnologías disruptivas a convertirse en una posible amenaza si su uso continúa evolucionando dentro de escenarios ofensivos.

En esta fase se esquematizaron los resultados, en tablas que muestran las características de cada estudio revisado abordándolo desde los objetivos que tiene cada uno de ellos y la relación o aporte al desarrollo de esta temática. Se encontró que varios estudios tienen el objetivo de analiza los modelos LLM basados en la arquitectura GPT-2, al ser un modelo de código abierto permite un control total y privacidad en el uso, para las investigaciones lo convierte, demostrando resultados excelentes en la generación de ataques sintéticos.(Ćirković et al., 2025), también se revisó estudios que objetivamente analizaron comparativas respecto a la automatización ofenseva en un ciclo completo de ataque. (Anis & Hammoudeh, 2025).

Si bien el enfoque de este estudio es el uso de modelos NLP y LLM en ciberataques, cada estudio presenta su enfoque en distintos ciberataques, en ataques phishing los LLM ha influido significativamente en su generación, que han demostrado que pueden ser muy eficaces (Sivaneswaran et al., 2026), existe un potencial del uso de LLM para potenciar loa ataques de escalada de privilegios (Happe et al., 2026), si bien se tienen estudios que abarcan en ellos varias pruebas de ciberataques similares a otros se ha integrado estudios con diferenciadores interesantes que podrían ser usados con facilidad como el impacto del quishing en correos maliciosos phishing que pueden eludir varios filtros (Weinz et al., 2025).

Tabla 3

Artículos revisados

Autor(es)	Título del estudio	Objetivo principal	Metodología	Principales hallazgos	Aporte al estudio
Bolaños Jiménez (2025)	Artificial Intelligence and Cybercrime: Contemporary Cyber Threats	Analizar las amenazas cibernéticas contemporáneas potenciadas por inteligencia artificial y los riesgos emergentes asociados al uso malicioso de estas tecnologías.	Investigación documental de carácter descriptivo basada en revisión y análisis de literatura especializada sobre IA y cibercrimen.	Se identifica usos ofensivos en phishing automatizado mediante WormGPT, generación de malware, robo de credenciales, deepfakes, creación de identidades falsas, ataques DDoS y automatización de distintas fases del ciclo de ataque.	Aporta evidencia sobre cómo las tecnologías basadas en IA, NLP y LLM están facilitando actividades ofensivas mediante automatización, personalización y escalabilidad de ataques.
Sivaneswaran et al. (2026)	A systematic literature review of large language models in phishing attack generation and detection	Analizar el uso de modelos LLM en la generación y detección de ataques phishing, evaluando su impacto en la automatización y efectividad de los ataques.	Revisión sistemática y análisis comparativo de investigaciones sobre phishing basado en LLM.	Los modelos GPT, LLaMA, Gemini y Claude permiten generar correos phishing muy convincentes, automatizan la creación sitios web maliciosos y reduciendo la barrera técnica para actores con conocimientos limitados. Adicionalmente se encuentra que los LLM pueden mejorar la evasión de sistemas de detección tradicionales mediante reescritura y personalización avanzada de contenido malicioso.	Se evidencia cómo los LLM y las técnicas NLP facilitan ataques ofensivos automatizados, incrementando la accesibilidad, escalabilidad y efectividad de campañas phishing ejecutadas por de manera fácil.
Webb et al. (2024)	Synthetic Social Engineering Scenario	Evaluar la capacidad de diferentes modelos de lenguaje	Investigación experimental mediante ajuste fino (fine-tuning)	Los modelos GPT-3.5 y especialmente Llama 3.1 fueron capaces de generar escenarios realistas de phishing, pretexting, spear	Se evidencia que los LLM pueden automatizar la generación de escenarios avanzados de ingeniería social y

	Generation Using LLMs for Awareness-Based Attack Resilience	para generar escenarios realistas de ingeniería social destinados a fortalecer la concienciación y resiliencia frente a ataques.	y prompt engineering sobre modelos BERT, T5, GPT-3.5 y Llama 3.1, utilizando escenarios de ingeniería social extraídos y procesados de *The Art of Deception*..	phishing, suplantación de identidad y otras técnicas de ingeniería social. Llama 3.1 Interactive tiene los mejores resultados en calidad, realismo, detalle y coherencia, demostrando una elevada capacidad para recrear ataques complejos mediante lenguaje natural.	phishing, reproduciendo tácticas ofensivas de manera convincente. Demuestra el potencial de estas tecnologías para reducir la complejidad de diseñar ataques basados en manipulación humana.
Wondimu et al. (2025)	Enabling Real-Time, Explainable DDoS Mitigation via On-Premise Large Language Models and Flow Analysis	Desarrollar un framework de detección y mitigación de ataques DDoS mediante Deep Learning y LLM locales para generar explicaciones y reglas automáticas de mitigación.	Investigación experimental utilizando el dataset CIC-IoT 2023, clasificación mediante YaTC e integración del modelo DeepSeek-7B para explicación y mitigación automática.	El modelo DeepSeek alcanzó altos niveles de precisión en detección de ataques DDoS como HTTP Flood, SYN Flood y TCP Flood con F1-score de 0.99. DeepSeek generó explicaciones comprensibles y reglas automáticas de mitigación con una reducción de tráfico malicioso superior al 90%.	Se evidencia cómo los LLM pueden integrarse en herramientas avanzadas de ciberseguridad para automatizar procesos de análisis y respuesta ofensiva/defensiva, demostrando la capacidad de estas tecnologías para reducir complejidad operativa y aumentar la automatización en ciberataques y mitigación de los mismo.
El Yagouby et al. (2025)	LLM-CVX: A Benchmarking Framework for Assessing the Offensive Potential of LLMs in Exploiting CVEs	Evaluar el potencial ofensivo de distintos LLM en la explotación automatizada de vulnerabilidades CVE mediante herramientas como	Investigación experimental y benchmarking sobre 14 modelos LLM (GPT, Claude, Gemini, DeepSeek, LLaMA, entre otros), utilizando explotación automatizada de CVEs	Se muestran resultados demostraron que tanto modelos open-source como closed-source lograron explotar CVEs exitosamente utilizando exploits públicos y herramientas ofensivas automatizadas. GPT4.1 obtuvo los mejores resultados de efectividad. El estudio concluye que los LLM pueden asistir o automatizar tareas ofensivas, reducir la barrera técnica y aumentar la accesibilidad de ataques	Se evidencia directamente cómo los LLM pueden facilitar actividades ofensivas de ciberseguridad mediante automatización exploits a vulnerabilidades, incrementando la escalabilidad y reduciendo la complejidad técnica requerida para ejecutar ataques.

		Metasploit y GitHub PoCs.	bajo el framework LLM-CVX	automatizados para actores con conocimientos básicos.	
Alqahtani et al. (2026)	Large Language Models for Cybersecurity Intelligence: A Systematic Review of Emerging Threats, Defensive Capabilities, and Security Evaluation Frameworks	Analizar el impacto de los LLM en la ciberseguridad, abordando amenazas ofensivas, capacidades defensivas y frameworks de evaluación de seguridad.	Revisión sistemática guiada por metodología PRISMA, sintetizando 167 estudios revisados por pares publicados entre 2022 y 2025.	El estudio concluye que los LLM incrementan significativamente la escalabilidad, adaptabilidad y realismo de ataques ofensivos como phishing automatizado, generación de malware, exploits y ataques de ingeniería social, reduciendo notablemente la barrera técnica para realizar ataques sofisticados. Además, identifica que la innovación ofensiva avanza más rápido que las capacidades defensivas y los mecanismos de gobernanza.	Se evidencia de manera directa cómo las tecnologías basadas en NLP y LLM pueden facilitar actividades ofensivas y automatizar procesos de ciberataque, permitiendo que actores con conocimientos limitados ejecuten acciones complejas mediante IA generativa. Respalda también el enfoque de concientización sobre la tendencia de estas tecnologías disruptivas a convertirse en amenazas emergentes.
Stefan et al. (2025)	Utilizing Fine-Tuning of Large Language Models for Generating Synthetic Payloads: Enhancing Web Application Cybersecurity through	Desarrollar y evaluar un modelo LLM ajustado mediante fine-tuning para generar payloads ofensivos automatizados orientados a pruebas de penetración en aplicaciones web.	Investigación experimental utilizando GPT-2 ajustado mediante fine-tuning sobre un dataset de 199,797 registros relacionados con ataques XSS, SQL Injection y Command Injection. Se implementó una aplicación web basada	El estudio demostró que los LLM permiten automatizar la generación de payloads ofensivos sofisticados para ataques XSS, SQL Injection y Command Injection. Los payloads generados fueron funcionales en entornos DVWA, evidenciando que el modelo puede generar ataques efectivos y reducir significativamente el tiempo y complejidad de preparación de pruebas ofensivas. La investigación concluye que el uso de LLM mejora la cobertura y precisión de pruebas de	Se evidencia directamente como los LLM pueden facilitar la automatización de actividades ofensivas en ciberseguridad mediante generación de payloads maliciosos y pruebas de penetración automatizadas. El estudio demuestra que herramientas basadas en NLP y LLM reducen la barrera técnica necesaria para ejecutar ataques complejos.

	Innovative Penetration Testing Techniques		en Flask y Raspberry Pi 5 para generar payloads ofensivos en tiempo real.	vulnerabilidades respecto a métodos tradicionales.	
Jones et al. (2025)	Analysing the Role of LLMs in Cybersecurity Incident Management	Analizar el papel de los modelos LLM dentro de la gestión de incidentes de ciberseguridad y evaluar su impacto en procesos ofensivos y defensivos relacionados con automatización y análisis de amenazas.	Investigación analítica y revisión técnica enfocada en el uso de LLM dentro de escenarios de gestión de incidentes y ciberseguridad.	El estudio muestra que los LLM están transformando la gestión de incidentes mediante automatización de análisis, generación contextual de información y apoyo en tareas relacionadas con detección y respuesta ante amenazas. También plantea preocupaciones relacionadas con el potencial uso ofensivo de estas tecnologías por actores maliciosos para automatizar procesos de ataque y reducir la complejidad operativa necesaria.	El estudio aporta evidencia conceptual sobre cómo los LLM pueden influir tanto en escenarios defensivos como ofensivos de ciberseguridad. El estudio respalda la hipótesis de que estas tecnologías disruptivas pueden facilitar actividades ofensivas mediante automatización y asistencia contextual, reduciendo la necesidad de conocimientos técnicos avanzados y dando una mejor accesibilidad de ciertas capacidades de ciberataque.
Albarrak et al. (2026)	Natural Language Processing (NLP)-Based Frameworks for Cyber Threat Intelligence and Early Prediction of Cyberattacks in Industry 4.0: A Systematic	Analizar el uso de frameworks basados en NLP y LLM para inteligencia de amenazas cibernéticas y predicción temprana de ciberataques dentro de entornos Industry 4.0.	Revisión sistemática de literatura enfocada en aplicaciones NLP y LLM orientadas a detección de amenazas, automatización de análisis y predicción de ciberataques en entornos industriales inteligentes.	El estudio identifica que los modelos NLP y LLM incrementan significativamente la capacidad de automatización en análisis de amenazas, clasificación de incidentes y generación contextual de respuestas frente a ciberataques. Y reconoce que estas tecnologías poseen un carácter dual, ya que pueden ser utilizadas tanto para fortalecer defensas como para facilitar ataques automatizados, phishing avanzado, generación de malware y evasión de controles tradicionales.	Aporta evidencia sobre cómo las tecnologías basadas en NLP y LLM reducen la complejidad operativa de procesos ofensivos en ciberseguridad, permitiendo automatización, contextualización y escalabilidad de ataques.

	Literature Review				
Landeta et al. (2026)	LLMs applied to web scraping and web crawling: a systematic review	Analizar el impacto y evolución del uso de LLMs en técnicas de web scraping y web crawling, identificando tendencias, desafíos, herramientas y aplicaciones en distintos dominios tecnológicos.	Revisión Sistemática de Literatura (SLR) basada en metodología PRISMA sobre 91 estudios relacionados con web scraping y crawling utilizando LLM y NLP. El análisis incluyó herramientas, modelos utilizados, métricas de evaluación, tendencias evolutivas y aplicaciones en múltiples sectores.	El estudio evidencia una transición significativa desde métodos tradicionales basados en reglas y arquitecturas semánticas impulsadas por LLM. El 94.5% de los trabajos analizados utilizaron modelos transformer-based como GPT, BERT o LLaMA, mientras que solamente el 5.5% mantuvo enfoques tradicionales basados en regex o selectores DOM. Se identificó que los LLM permiten automatizar generación de código, razonamiento contextual, navegación autónoma, generación de XPaths y adaptación dinámica frente a cambios en interfaces web. También se destaca su uso en monitoreo de dark web, OSINT, threat intelligence y detección de phishing	Muestra evidencia empírica sólida sobre cómo los LLM están reduciendo la complejidad técnica de actividades relacionadas con automatización avanzada, extracción de información y análisis ofensivo dentro de escenarios de ciberseguridad y OSINT. Demuestra que estas tecnologías permiten automatizar tareas complejas mediante generación contextual y razonamiento autónomo, incrementando la accesibilidad de capacidades técnicas avanzadas incluso para actores con conocimientos limitados.
Anis et al. (2025)	Weaponizing AI in Cyberattacks: A Comparative Study of AI-powered Tools for Offensive Security	Analizar cómo herramientas impulsadas por IA pueden automatizar distintas fases de la ciberseguridad ofensiva y evaluar comparativamente su efectividad frente a	Investigación experimental y comparativa utilizando herramientas automatizadas de offensive security como WebCopilot, RustScan y Nmap dentro del modelo Cyber Kill Chain	El estudio demuestra que la IA permite automatizar múltiples fases del Cyber Kill Chain mediante NLP y LLM para reconocimiento y phishing, Deep Learning para generación de exploits, Reinforcement Learning para adaptación ofensiva y GANs para evasión de detección. Los resultados experimentales evidenciaron que herramientas automatizadas como WebCopilot y RustScan	Aporta evidencia empírica sobre cómo tecnologías basadas en NLP, LLM y Machine Learning facilitan la automatización de actividades ofensivas en ciberseguridad. El estudio demuestra que herramientas impulsadas por IA reducen la complejidad técnica y el tiempo requerido para ejecutar tareas ofensivas como reconocimiento,

		herramientas tradicionales.	(CKC). El estudio integró técnicas NLP, ML, DL, RL y LLM para automatizar reconocimiento, weaponization, delivery y explotación de ataques	superaron significativamente a herramientas tradicionales en velocidad y efectividad de reconocimiento y escaneo de red. Se concluye que la automatización ofensiva basada en IA incrementa la escalabilidad y reduce considerablemente la intervención humana requerida en ciberataques.	phishing, escaneo y explotación, permitiendo que actores con menor experiencia técnica puedan desarrollar ataques más eficientes y escalables.
Yigit et al. (2025)	Review of Generative AI methods in cybersecurity	Analizar el impacto de la IA generativa (GenAI) en la ciberseguridad desde perspectivas ofensivas, defensivas, éticas y sociales, identificando riesgos, oportunidades y desafíos emergentes.	Revisión sistemática y análisis crítico de investigaciones recientes sobre GenAI en ciberseguridad, incluyendo evaluación de técnicas como jailbreak, prompt injection, phishing automatizado, generación de malware y hacking asistido por LLM.	El estudio muestra que los LLM pueden facilitar ataques de ingeniería social, phishing automatizado, generación de malware, creación de payloads, reconocimiento de objetivos y escalamiento de privilegios. Asimismo, demuestra que técnicas como jailbreak y prompt engineering pueden eludir mecanismos de seguridad de algunos modelos.	Muestra evidencia integral sobre cómo los LLM y tecnologías basadas en NLP pueden automatizar distintas fases de un ciberataque, desde la ingeniería social hasta la generación de malware y explotación de vulnerabilidades.
Weinz et al. (2025)	The Impact of Emerging Phishing Threats: Assessing Quishing and LLM-generated	Evaluar la efectividad de correos phishing generados mediante LLM y ataques Quishing (QR phishing) en entornos	Investigación experimental basada en simulaciones de phishing realizadas en tres organizaciones de distinto tamaño, utilizando correos	Los correos generados mediante LLM demostraron ser altamente efectivos, especialmente en organizaciones pequeñas y medianas. Los investigadores concluyen que la combinación de información OSINT con LLM permite generar campañas phishing personalizadas, de bajo costo y con alta	Aporta evidencia empírica de que los LLM pueden facilitar la creación de campañas de phishing altamente convincentes mediante lenguaje natural y datos públicos, reduciendo significativamente la complejidad técnica

	Phishing Emails against Organizations	organizacionales reales.	tradicionales, correos con códigos QR y correos generados por LLM alimentados con información OSINT	capacidad de engaño. Además, los correos fueron creados utilizando la versión gratuita de ChatGPT y datos públicos,	requerida para diseñar ataques de ingeniería social.
Isozaki et al. (2025)	Towards Automated Penetration Testing: Introducing LLM Benchmark, Analysis, and Improvements	Evaluar la capacidad de distintos LLM para ejecutar tareas de pruebas de penetración mediante un benchmark diseñado para medir reconocimiento, explotación, escalamiento de privilegios y técnicas generales de pentesting.	Investigación experimental mediante la construcción de un benchmark de pentesting basado en máquinas vulnerables de Hack The Box. Se evaluaron modelos como GPT-4o y Llama 3.1-405B en tareas de reconocimiento, explotación, análisis de código, fuerza bruta, escalamiento de privilegios e ingeniería social.	Los resultados que muestra el estudio indica que los LLM pueden asistir de forma efectiva en diversas fases del pentesting, incluyendo enumeración de servicios, identificación de vulnerabilidades, explotación, análisis de configuraciones y generación de pasos para comprometer sistemas. El estudio evidencia que los modelos son capaces de planificar tareas complejas mediante listas de acciones y razonamiento secuencial.	EL estudio aporta evidencia directa de que los LLM pueden actuar como asistentes de pruebas de penetración, automatizando procesos de reconocimiento, explotación y análisis técnico.
Happe et al. (2026)	LLMs as Hackers: Autonomous Linux Privilege Escalation Attacks	Evaluar la capacidad de modelos LLM para ejecutar de forma autónoma procesos de escalamiento de	Investigación experimental basada en un benchmark propio de escalamiento de privilegios en Linux compuesto por múltiples	Los resultados del estudio evidenciaron que GPT-4 Turbo logró comprometer hasta el 66% de los escenarios de escalamiento de privilegios sin asistencia y hasta el 83% cuando recibió orientación de alto nivel, acercándose al desempeño de un pentester humano que	El estudio aporta evidencia empírica directa sobre la capacidad de los LLM para automatizar actividades ofensivas avanzadas relacionadas con pruebas de penetración y escalamiento de privilegios. Los hallazgos demuestran que estas

		privilegios en sistemas Linux mediante identificación y explotación de vulnerabilidades reales.	clases de vulnerabilidades (SUID, sudo, Docker, cron jobs, divulgación de credenciales y SSH keys). Se comparó el desempeño de GPT-4 Turbo, GPT-3.5 Turbo y Llama 3 frente a un pentester profesional.	alcanzó el 75% sin ayuda y 92% con asistencia. GPT-3.5 obtuvo tasas menores, pero también fue capaz de ejecutar ataques exitosos. El estudio concluye que los LLM modernos pueden identificar vulnerabilidades, interpretar resultados de enumeración y generar acciones de explotación de forma autónoma en múltiples escenarios Linux.	tecnologías pueden reducir significativamente la necesidad de conocimientos técnicos especializados durante la fase de explotación, permitiendo que procesos complejos de reconocimiento y ataque sean ejecutados mediante asistencia basada en lenguaje natural.
Czaja et al. (2025)	Cybersecurity challenges and opportunities of machine learning-based artificial intelligence	Analizar las oportunidades y riesgos que introduce la inteligencia artificial dentro de la ciberseguridad, considerando tanto aplicaciones defensivas como ofensivas.	Revisión de literatura sobre aplicaciones de IA y Machine Learning en múltiples dominios de ciberseguridad, incluyendo phishing, detección de intrusiones, deepfakes, IoT, cloud security, pentesting y autenticación.	El estudio evidencio que la IA posee una naturaleza dual dentro de la ciberseguridad. Por una parte, mejora capacidades de detección de amenazas, autenticación, análisis de malware y protección de infraestructuras. Por otra, facilita nuevas amenazas como deepfakes, ataques de envenenamiento de datos, automatización de ataques, explotación de vulnerabilidades y pruebas de penetración autónomas.	Aporta una visión integral sobre el papel dual de la IA en la ciberseguridad y respalda la hipótesis de que tecnologías basadas en Machine Learning, NLP y modelos avanzados de IA están reduciendo la complejidad de diversas actividades ofensivas. El estudio evidencia cómo la automatización impulsada por IA puede facilitar tareas de reconocimiento, explotación y pruebas de penetración.

Fase 3: Síntesis

En la matriz realizada en el análisis de la fase 2, se observa que el phishing es el ataque más estudiado en los artículos analizados. También se puede indicar que los modelos GPT, LLaMA y Gemini, al igual que son los más estudiados, también son los más usados.

Los artículos seleccionados debían ser actuales; se consideró que los estudios fuesen a partir de 2023, pero en el transcurso del desarrollo de la matriz se llegó a manejar artículos de 2025 a 2026, permitiendo que este estudio sea moderno referente al uso de LLM en tareas ofensivas.

En los estudios analizados, el patrón que se identifica es la reducción de la barrera técnica con el uso de los LLM en tareas de ciberseguridad y ciberataque, enfocándose especialmente en este último, que es el objetivo del estudio. Los LLM, en todos los estudios analizados, destacan por automatizar las tareas ofensivas de ataque, como explotación, pentesting, phishing y DDoS. Además, permite generar un mayor realismo en ataques relacionados con la ingeniería social y, concluyendo en gran parte de los artículos, la tendencia hacia una mayor escalabilidad de estas capacidades

Los LLM permite genera mensajes altamente personalizados y convincentes, en estudios realizados sobres arquitecturas BERT, T, GPT-3.5 y LLAMA 3.1 se demuestra la capacidad de uso de estas arquitecturas para generar escenarios de ingeniero social de alta calidad, delincuentes pueden hacer emplear estas herramientas para diseñar ataques dirigiros contra personas u organizaciones.(Webb et al., 2025), demostrando la alta efectividad de ataques phishing y spear phishing con agregados de quishing.

La generación de malware y payloads mediante modelos LLM permite una reducción considerable del tiempo de para el desarrollo de ataques. El acceso que pueden tener de los atacantes a herramientas más sofisticadas de inteligencia artificia generativa suscita una preocupación sobre la posibilidad de construir ataques más avanzados (Yigit et al., 2025), desde una perspectiva ética, hay desafíos que varios estudios plantearon abordar enfocados en evitar el uso indebido de los modelos LLM y NLP.

Uno de los vectores de ataque ofensivo es la explotación de vulnerabilidades en este contexto el uso de las LLMs tanto modelos propietarios como de código abiertos muestra una capacidad ofensiva exitosa y eficiente para explotar vulnerabilidades CVE, haciendo uso de herramientas públicas de explotación automatizando todas las tareas de ejecución.(Isozaki et al., 2025). La automatización y eficiencia que ofrecen estas herramientas hacen que el uso de LLMs destaque por su desempeño ofensivo.

La etapa de reconocimiento, es la etapa inicial donde un atacante realiza la recolección de información para planificar una intrusión, si bien no es un ataque ofensivo como tal representa parte de la ciber kill chain, el uso de LLMs como WebCopilot y Rust Scan muestra una ventaja en la automatización para la enumeración de dominios y subdominios, escaneo de redes superando herramientas específicamente creadas para ese objetivo como Sublist3r y Nmap,

haciendo escaneos con mayor rapidez, donde el tiempo y la exhaustividad son factores críticos. (Anis & Hammoudeh, 2025)

CONCLUSIONES

La revisión sistemática que se ha realizado en el presente estudio a identificado diversas herramientas basadas en Natural Language Processing (NLP), Large Language Models (LLM) y técnicas de Machine Learning están facilitando la ejecución de actividades ofensivas relacionados a ciberataques. Se han analizado varias herramientas que se implican en varias partes de Ciber Kill Chain y que han demostrado rapidez, facilidad de uso y aportando una alta efectividad.

P1: ¿Los LLMs está reduciendo la complejidad de ejecutar ciberataques para personas con poco conocimiento técnico?

Las conclusiones y revisiones de cada articulo permite afirmar que sí, los estudios analizados coinciden dentro de sus conclusiones y de sus resultados que los LLM y NLP están disminuyendo significativamente la complejidad técnica y operativa necesaria para ejecutar actividades ofensivas. Diversas investigaciones destacan incluso el uso de las herramientas en fases de reconocimiento y enumeración de objetivos, explotación de vulnerabilidades y automatización de tareas que antes eran complejas. La facilidad que se le da al uso de LLMs es uno de los principales factores identificados, puesto que las instrucciones son expresadas en lenguaje natural e interpretadas por estas herramientas en a acciones técnicas concretas.

Estos hallazgos sugieren un cambio revolucionario en el perfil tradicional de un actor maliciosos, si anteriormente se requería experiencia avanzada en programación, redes, conocimiento de sistemas operativos y seguridad informática para realizar ataques avanzados, ahora pueden realizados con conocimientos básicos limitados apalancándose en el uso de herramientas LLM. Como consecuencia la curva de aprendizaje disminuye y la barrear para desarrollar actividades ofensivas disminuye, incrementando la accesibilidad a capacidades que anteriormente estaban reservadas a usuarios con formación especializada.

P2: ¿Qué herramientas basadas en Natural Language Processing (NLP) y Large Language Models (LLM) o en funciones de Machine Learning están facilitando ataques a actores con conocimientos limitados?

Los resultados destacan herramientas basadas en modelos como GPT, LLaMA, Gemini, Claude y DeepSeek, las cuales son utilizadas de forma directa mediante integraciones con plataformas y herramientas de automatización. La literatura analizada muestra que estas tecnologías se emplean principalmente en ataques de phishing, ingeniería social, generación de malware, creación de payloads, explotación de vulnerabilidades, escalamiento de privilegios y actividades de reconocimiento inicial.

En conjunto, los estudios coinciden en que estas herramientas permiten automatizar procesos que anteriormente requerían experiencia especializada, transformando instrucciones de

lenguaje natural en acciones técnicas avanzadas. También varios estudios destacan su utilización como asistentes en pentesting, automatización de etapas de reconocimiento y enumeración, así como en creacion de plataformas diseñadas para facilitar actividades ofensivas.

Al analizar estos hallazgos bajo el concepto de Cyber Kill Chain, se observa que el uso de estas herramientas si bien se reconoce con mayor presencia en las etapas de Armamento, entrega y explotación que son fases donde la el conocimiento técnico es importante, también la herramienta se hace uso en las fases iniciales del ataque, particularmente en reconocimiento, haciendo que esta se realice de forma automática y con mayor eficiencia que otras herramientas creadas específicamente para dicha función.

P3: ¿Hasta qué punto la efectividad de uso de LLM y NLP en estas herramientas puede ser aprovechada para fines maliciosos?

La revisión de los artículos permitió concretar que la efectividad del uso de la inteligencia artificial esa provechado en un alto grado para fines maliciosos. Los estudios en el uso de LLM y NLP muestran las capacidades de automatización, personalización, la generación de contenido manteniendo y análisis contextual, facilitando las capacidades de producir correos phishing altamente convincentes, generar scripts maliciosos, la asistencia en la explotación de vulnerabilidades y la automatización de las fases de reconocimientos y recopilación de información de objetivos de ataques.

La capacidad de estas herramientas de comprender el lenguaje natural permite una interacción sencilla con los actores maliciosos que poseen poco conocimiento técnico proporcionando de esta manera una asistencia que tradicionalmente requiere una formación más especializada. Esta situación incrementa la escalabilidad de los ataques, reduce los costos operativos para los actores maliciosos y permite la ejecución de campañas más personalizadas y difíciles de detectar.

No obstante, es importante señalar que gran parte de los estudios revisados fueron desarrollados en entornos controlados o en situaciones experimentales. Es por esto que aunque los resultados evidencian un potencial significativo para la automatización de actividades ofensivas, existe la necesidad de continuar investigando el comportamiento de estas herramientas en escenarios reales y frente a mecanismos defensivos más avanzados que incluyan IA y modelos de machine learning para detectar nuevas amenazas.

REFERENCIAS

Albarrak, M., Salonitis, K., & Jagtap, S. (2026). Natural Language Processing (NLP)-Based Frameworks for Cyber Threat Intelligence and Early Prediction of Cyberattacks in Industry 4.0: A Systematic Literature Review. *Applied Sciences, 16*(2), 619. https://doi.org/10.3390/app16020619

Alqahtani, H., & Kumar, G. (2026). Large Language Models for Cybersecurity Intelligence: A Systematic Review of Emerging Threats, Defensive Capabilities, and Security Evaluation Frameworks. *Computers, Materials & Continua, 87*(3). https://doi.org/10.32604/cmc.2026.077367

Anis, F., & Hammoudeh, M. (2025). Weaponizing AI in Cyberattacks A Comparative Study of AI powered Tools for Offensive Security. *Proceedings of the 8th International Conference on Future Networks & Distributed Systems, ICFNDS '24*, 283–290. https://doi.org/10.1145/3726122.3726164

Ćirković, S., Mladenović, V., Tomić, S., Drljača, D., & Ristić, O. (2025). Utilizing Fine-Tuning of Large Language Models for Generating Synthetic Payloads: Enhancing Web Application Cybersecurity through Innovative Penetration Testing Techniques. *Computers, Materials & Continua, 82*(3), 4409–4430. https://doi.org/10.32604/cmc.2025.059696

Czaja, P., Gdowski, B., Niemiec, M., Mees, W., Stoianov, N., Votis, K., Kharchenko, V., Katos, V., & Merialdo, M. (2025). Cybersecurity challenges and opportunities of machine learning-based artificial intelligence. *Neural Computing and Applications, 37*(33), 27931–27956. https://doi.org/10.1007/s00521-025-11604-9

El yagouby, M. A., Lahmadi, A., Zakroum, M., Festor, O., & Ghogho, M. (2026). LLM-CVX: A Benchmarking Framework for Assessing the Offensive Potential of LLMs in Exploiting CVEs. *Proceedings of the 18th ACM Workshop on Artificial Intelligence and Security, AISec '25*, 194–205. https://doi.org/10.1145/3733799.3762978

Happe, A., Kaplan, A., & Cito, J. (2026). LLMs as Hackers: Autonomous Linux Privilege Escalation Attacks. *Empirical Software Engineering, 31*(3), 70. https://doi.org/10.1007/s10664-025-10758-3

Isozaki, I., Shrestha, M., Console, R., & Kim, E. (2025). Towards Automated Penetration Testing: Introducing LLM Benchmark, Analysis, and Improvements. *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization, UMAP Adjunct '25*, 404–419. https://doi.org/10.1145/3708319.3733804

Jiménez, C. B. (2025). Inteligencia Artificial y Cibercrimen: Amenazas cibernéticas contemporáneas. *Estudios en Seguridad y Defensa, 20*(40), 199–220. https://doi.org/10.25062/1900-8325.5074

Jones, G., Kasimatis, D., Pitropakis, N., Macfarlane, R., & Buchanan, W. J. (2025). Analysing the role of LLMs in cybersecurity incident management. *International Journal of Information Security, 24*(6), 228. https://doi.org/10.1007/s10207-025-01144-7

Landeta-López, P., García, J. M., Guevara-Vega, C., & Ruiz-Cortés, A. (2026). LLMs applied to web scraping and web crawling: A systematic review. *Computing, 108*(6), 78. https://doi.org/10.1007/s00607-026-01666-5

Sivaneswaran, D., Hewage, C. T. E. R., Herath, H. M. K. K. M. B., Rathore, R. S., Singh, V. K., & Jiang, W. (2026). A systematic literature review of large language models in phishing attack generation and detection. *Array, 30*, 100775. https://doi.org/10.1016/j.array.2026.100775

Webb, J., Abri, F., & Akther, S. (2025). Synthetic Social Engineering Scenario Generation Using LLMs for Awareness-Based Attack Resilience. *IEEE Access, 13*, 174831–174856. https://doi.org/10.1109/ACCESS.2025.3614550

Weinz, M., Zannone, N., Allodi, L., & Apruzzese, G. (2025). The Impact of Emerging Phishing Threats: Assessing Quishing and LLM-generated Phishing Emails against Organizations. *Proceedings of the 20th ACM Asia Conference on Computer and Communications Security, ASIA CCS '25*, 1550–1566. https://doi.org/10.1145/3708821.3736195

Wondimu, H., Alfatemi, A., Rahouti, M., Chehri, A., Weichbroth, P., & Ghani, N. (2025). Enabling Real-Time, Explainable DDoS Mitigation via On-Premise Large Language Models and Flow Analysis. *Procedia Computer Science, 270*, 1390–1399. https://doi.org/10.1016/j.procs.2025.09.260

Yigit, Y., Buchanan, W. J., Tehrani, M. G., & Maglaras, L. (2025). Review of Generative AI methods in cybersecurity. *Internet of Things and Cyber-Physical Systems, 5*, 241–261. https://doi.org/10.1016/j.iotcps.2026.04.001