The impact of artificial intelligence on enterprise pentesting: an analysis through social engineering and generative tools

Impacto de la inteligencia artificial en el pentesting empresarial, análisis mediante técnicas de ingeniería social y herramientas generativas

Autores:

Barreiro-Molina, Franklin David
UNIVERSIDAD CATOLICA DE CUENCA
Unidad Académica de Informática, Ciencias de la Computación, e Innovación Tecnológica
Cuenca-Ecuador

franklin.barreiro.87@est.ucacue.edu.ec

https://orcid.org/0009-0004-2149-5652

Cuenca-Tapia, Juan Pablo
UNIVERSIDAD CATOLICA DE CUENCA
Unidad Académica de Informática, Ciencias de la Computación, e Innovación Tecnológica

jcuenca@ucacue.edu.ec

https://orcid.org/0000-0001-5982-634X

Fechas de recepción: 03-AGO-2025 aceptación: 03-SEP-2025 publicación: 30-SEP-2025

https://orcid.org/0000-0002-8695-5005
http://mqrinvestigar.com/

Resumen

La creciente sofisticación de los ciberataques y sus variantes de ingeniería social y ransomware ha hecho imperativo revisar las defensas cibernéticas organizacionales. Por lo cual esta investigación investiga el efecto del uso de herramientas de inteligencia artificial generativa, como ChatGPT en pruebas de penetración centradas en explotar vulnerabilidades humanas, este estudio de caso se llevó a cabo con la empresa Casystem situada en Cuenca, Ecuador mediante la comparación de dos grupos de campañas de phishing simuladas una tradicional y una asistida por inteligencia artificial. Se aplicó un enfoque de método mixto con análisis cuantitativo y cualitativo centrado en la tasa de clics, la presentación de credenciales el tiempo de respuesta y los métricas de índice de riesgo humano (IRH), los resultados muestran que en la campaña asistida por inteligencia artificial el índice de riesgo humano (IRH) fue del 53.3% mientras que el enfoque tradicional arrojó un 13.3% lo que ilustra que el contenido generado por IA fue mucho más persuasivo, también se observaron riesgos éticos y legales derivados del marco corporativo no regulado en el uso de estas herramientas.

Palabras clave: Ciberseguridad; pentesting; ingeniería social; inteligencia artificial generativa; phishing; ransomware

Abstract

The increasing sophistication of cyberattacks and their social engineering and ransomware variants has made it imperative to review organizational cyber defenses. Therefore, this research investigates the effect of using generative artificial intelligence tools, such as ChatGPT, in penetration tests focused on exploiting human vulnerabilities. This case study was conducted with the company Casystem, located in Cuenca, Ecuador, by comparing two groups of simulated phishing campaigns: one traditional and one assisted by artificial intelligence. A mixed-method approach was applied with quantitative and qualitative analysis focusing on click-through rate, credential presentation, response time, and human risk index (HRI) metrics. The results show that in the AI-assisted campaign the human risk index (HRI) was 53.3% while the traditional approach yielded 13.3% illustrating that AI-generated content was much more persuasive. Ethical and legal risks were also observed arising from the unregulated corporate framework in the use of these tools.

Keywords: Cybersecurity; pentesting; social engineering; generative artificial intelligence; phishing; ransomware

Introducción

En la actualidad, la conectividad, la automatización y la transformación digital impulsan todos los sectores productivos. Esa conveniencia incluye un aumento en las amenazas cibernéticas que dificultan mantener la seguridad de los datos dentro de las organizaciones. En Cuenca, Ecuador, la tercera ciudad más grande del país, se han reportado casos realmente significativos de ciberataques. Solo en 2023, se reportaron más de 2,400 incidentes de seguridad al Equipo de Respuesta a Emergencias Informáticas de Ecuador (EcuCERT), lo que causó un impacto económico y operativo considerable mediante ataques de ransomware y técnicas de ingeniería social.

De acuerdo con lo establecido por López y Zatarain (2023) "la ingeniería social se define como "un conjunto de técnicas que vulneran a las personas con el fin de obtener acceso a información o sistemas" y representa un de los métodos más efectivos para comprometer la seguridad organizacional, ya que "se trata del eslabón más débil del ser humano en cuanto a seguridad". Dentro de tal contexto, las pruebas de penetración conocida también como Pentesting evolucionaron de ser una mera referencia para convertirse en un recurso indispensable que ayuda a identificar amenazas antes de que sean explotadas. Imitando en cierta medida los ataques reales para verificar si los sistemas o las personas son resistentes a estos.

Sin embargo, el advenimiento de la inteligencia artificial generativa y ChatGPT en particular, ha añadido una nueva dimensión de peligro lo que hace a estas nuevas herramientas capaces de crear correos, scripts y contenidos de ciberseguridad completamente personalizados y realistas, aparentemente enviados por autores exclusivos. De acuerdo con Linares y Salazar, (2025), la IA generativa puede ser utilizada tanto para reforzar la seguridad como para potenciar los ataques cibernéticos lo que plantea un dilema ético importante sobre su aplicación en escenarios empresariales (Sacoto et al., 2025).

Además, investigaciones más recientes, como la de Santos (2024), manifiestan que el uso de GenAI en pruebas de pentesting puede aumentar de forma significativa el éxito en ataques

simulados de phishing al permitir que el evaluador con antela los fallos humanos que pueden ser corregidos antes de que intentos de hackeo puedan realizarse. Sin embargo, la falta de normativas puede ocasionar el uso de estas tecnologías de una forma que atente la privacidad o algún marco legal, particularmente en el manejo de información sensible sin el permiso expreso de los dueños.

Por lo tanto, este estudio se trata de descubrir cómo usar ataques de Ransomware de IA para falsificar lo que puede ayudar a las empresas a probar mejor sus defensas. Dentro del presente estudio Casystem, tienen una configuración sólida de 10 host para su respectiva utilización, motivo por el cual se compararán los trucos de phishing tradicionales con los que usan la IA para ver cuál es mejor para detectar cuando la gente se engaña.

Con respecto al riesgo específico de la IA el Marco de Gestión de Riesgos de IA del NIST (AI RMF 1.0) incluido su perfil para IA generativa ofrece un enfoque de gestión de riesgos a lo largo del ciclo de vida (gobernanza, mapeo, medición y gestión) que es útil para evaluar el sesgo, la trazabilidad, la seguridad y la gestión responsable de modelos dentro de contextos corporativos, además la integración de AI RMF y SGSI (ISO/IEC 27001) mejora las salvaguardias técnicas y organizativas para el uso seguro de asistentes generativos durante las pruebas internas (Scarfone et al., 2021).

Ademas, NIST SP 800-115 proporciona una guía documentada para la planificación y ejecución de pruebas de seguridad, incluyendo pruebas de penetración, análisis de hallazgos y formulación de estrategias de mitigación. Su adopción ayuda a estructurar campañas de ingeniería social controladas, ya sean tradicionales o asistidas por IA, dentro de un marco equilibrado que mantiene el consentimiento informado, el perímetro designado y salvaguardas para el entorno de producción del organismo (Fitriana et al., 2023). Complementariamente, NIST SP 800-53 Rev. 5 cataloga controles de seguridad y privacidad que son mapeables a salvaguardas organizacionales contra campañas de phishing impulsadas por IA.

Evidencia internacional sobre IA y campañas de ingeniería social

Informes recientes coinciden con la afirmación de que la IA está aumentando la sofisticación y la tasa de éxito de los ataques de ingeniería social. El informe de (ENISA, 2022) sobre el panorama de amenazas 2024 indica que la IA generativa normaliza la creación de texto, audio y video convincentes lo que permite campañas de phishing y deepfake altamente personalizadas. Para el sector financiero europeo, ENISA, (2022) informa sobre impactos económicos directos de las campañas de phishing habilitadas por IA que integran Ai as a Tool con tácticas AiTM y BEC.

Las autoridades nacionales también han emitido diagnósticos sobre este fenómeno. Según la NCSC, (2021) del Reino Unido, la IA incrementará el volumen y el impacto de las estafas por correo y ransomware, a la vez que dificultará su detección por los usuarios y por los sistemas de defensa convencionales. Los recientes casos de herramientas de IA para el crimen, como "GhostGPT", muestran la existencia de modelos peligrosos que no poseen salvaguardas y pueden generar, de manera industrial, correos, phishing y código dañino.

Al momento de realizar una investigación no solo hay que pensar en qué tan bien funcionan estas herramientas, sino también en las herramientas que son inteligentes, morales y correctas. En base a la conclusión que se obtuvo en dicha investigación, se espera crear consejos que se puedan poner en práctica una y otra vez, que en la ciberseguridad la mejorarán en ecosistemas análogos y, reitero, dentro del marco normativo LOPDP de Ecuador.

Objetivos de la investigación

A medida que los ataques cibernéticos continúan creciendo y afectan gravemente a las organizaciones mediante ransomware las pruebas de penetración están adquiriendo un rol crucial para prevenir y mitigar incidentes la irrupción de la Inteligencia Artificial generativa y su aplicación a sistemas como ChatGPT pone de manifiesto un potencial significativo tanto para reforzar como para vulnerar los sistemas de seguridad especialmente a través de técnicas

de ingeniería social. Por lo que la presente investigación se plantea analizar cómo pueden utilizarse estas tecnologías emergentes de manera controlada y ética para robustecer los sistemas de protección de la información en un contexto corporativo en entornos de trabajo reales.

Objetivo general

Evaluar el impacto de la Inteligencia Artificial en la ejecución de pruebas de penetración empresarial mediante técnicas de ingeniería social con el fin de fortalecer la prevención, detección y respuesta de ataques cibernéticos como el ransomware fortaleciendo así la seguridad de la información en las organizaciones.

Objetivos específicos

1. Caracterizar las herramientas y soluciones actuales basadas en Inteligencia Artificial utilizadas específicamente en pruebas de pentesting con ingeniería social.
2. Analizar beneficios operativos, técnicos y estratégicos que aportan estas herramientas.
3. Evaluar mediante un caso práctico, el impacto específico que tiene la implementación de estas herramientas en la identificación y mitigación de vulnerabilidades asociadas a ataques de ransomware.
4. Examinar los riesgos éticos y legales derivados del uso de Inteligencia Artificial generativa en las pruebas internas de pentesting mediante técnicas de ingeniería social, considerando aspectos normativos.

Justificación

El aumento de la dependencia tecnológica y el avance de la transformación digital de los negocios en Cuenca han convertido la seguridad de la información en una dimensión estratégica para salvaguardar la continuidad del negocio y proteger los activos más

importantes. El ransomware y las tácticas de ingeniería social constituyen algunos de los ataques más comunes y peligrosos. Estas técnicas apuntan a las debilidades humanas o se camuflan a través de falsos pretextos para alcanzar una zona crítica e interrumpir los servicios, incitar acusaciones legales, arruinar la imagen, etc. Y no solo resultan en violaciones de datos, sino también en pérdidas monetarias inmensas y, en última instancia, en la tambaleante confianza del cliente, lo que lleva a organizaciones potencialmente insostenibles del mañana.

La implementación del estudio en una empresa ubicada en Cuenca documentara resultados importantes, con recomendaciones aplicables a otras empresas de características similares. Además, esta IA no solo podrá mecanizar las tareas de detección, sino también proporcionar recomendaciones concretas para una respuesta rápida y mitigación de riesgos (Pérez 2023), jugando a favor especialmente de empresas que manejan información altamente sensible (García et al. 2025).

Sin embargo, los temores sobre los peligros que plantean estas herramientas se reconocen en nuevas investigaciones que sugieren que el uso irresponsable y apresurado de herramientas como ChatGPT corre el riesgo de permitir a actores malintencionados construir malware de última generación explotable para ransomware y métodos de ataque informático dirigidos (Martínez 2024).

Alcance del estudio

Este estudio se desarrollará como un caso práctico en la compañía Casystem que cuenta con una infraestructura tecnológica compuesta por diez hosts el objetivo principal es analizar el impacto que tiene el uso de IA generativa en especial el uso de la herramienta ChatGPT en la automatización de procesos internos de pentesting que se llevan a cabo con técnicas de ingeniería social enfocadas en identificar posibles puntos de vulnerabilidad a ransomware en el ojo humano.

El trabajo de investigación incluye el diseño, implementación y análisis de campañas de phishing que se ejecutan por medio de escenarios realistas creados a través de Inteligencia

Artificial la actividad está planteada en un grupo controlado de colaboradores en un ambiente controlado y autorizado por la dirección de la empresa siempre garantizando la integridad de los datos personales y el respeto a la normativa vigente de privacidad como la Ley Orgánica de Protección de Datos Personales LOPDP.

La investigación propuesta no contempla llevar a cabo ataques técnicos a la infraestructura como redes o servidores ni incluye el análisis forense y la respuesta a incidentes post-ataque. Los conocimientos adquiridos permitirán evaluar el grado de exposición humana a ataques contemporáneos de ingeniería social así como a desarrollar estrategias prácticas y replicables para organizaciones en la misma industria o con vulnerabilidades similares, dado que se trata de un estudio contextual centrado en una sola empresa los resultados pueden diferir al ser extrapolados a otras organizaciones con diferentes estructuras o niveles de madurez en ciberseguridad.

El alcance de esta investigación incluye el desarrollo de charlas de capacitación y concienciación dirigidas al personal técnico y administrativo estas estarán orientadas al adiestramiento sobre identificación y respuesta a intentos de ingeniería social perpetrados por IA y en particular a ataques de ransomware. Igualmente se definirán roles y responsabilidades que garanticen el desarrollo ético y seguro de los ensayos.

Al final de la investigación se realizará la entrega de un reporte que sintetiza todos los resultados, los cuales incluyen la vulnerabilidad humana que se logró identificar en las pruebas asistidas por ChatGPT, además se agrega una evaluación comparativa de las ventajas técnicas, operativas y en materia de prevención en relación a los enfoques tradicionales, y finalmente se incluyen recomendaciones formuladas para la integración segura y efectiva de herramientas de Inteligencia Artificial generativa a los procesos internos de ciberseguridad.

Limitaciones del estudio

El estudio tiene limitaciones particulares que delinean la investigación de los hallazgos. Dado que el tamaño de la muestra fue de 15 participantes, generalizar estos resultados a otras organizaciones o sectores se vuelve altamente limitado. El hecho de que la selección no sea aleatoria añade una fuente potencial de sesgo y compromete cuán verdaderamente

representativos son sus datos. Estas pruebas se realizarán en un entorno controlado y autorizado, lo cual no siempre representa el comportamiento real de los usuarios bajo condiciones operativas. Cabe destacar que no hay aprendizaje sobre los resultados futuros de las campañas de concienciación y, por lo tanto, no hay información sobre si el entendimiento se extiende más allá de la campaña inmediata o si el cambio de conducta persiste.

Material y métodos

Material

La investigación se realizó mediante la realización de evaluaciones internas de ingeniería social en un entorno empresarial limitado con una base tecnológica, humana y metodológica predeterminada. Los principales medios para lograr esto fueron campañas de phishing que aprovecharon la vulnerabilidad humana utilizando la inteligencia artificial generativa ChatGPT para generar respuestas únicas. También se diseñaron encuestas de percepción, plantillas de documentación de incidentes y hojas de observación, que son algunas herramientas desarrolladas para la recopilación de datos cuantitativos y cualitativos. En donde los datos fueron sistematizados mediante hojas de cálculo, facilitando el análisis de métricas clave como tasa de clics, tiempo de reacción, porcentaje de usuarios comprometidos y nivel de riesgo percibido.

Cabe destacar que la totalidad del proceso experimental fue autorizado y ejecutado en la empresa Casystem, ubicada en Cuenca, que cuenta con una infraestructura tecnológica compuesta por diez hosts. Las pruebas fueron monitoreadas en condiciones seguras, garantizando la integridad de los sistemas, la confidencialidad de los datos y el consentimiento informado de los participantes.

Métodos

La organización de la investigación se llevó a cabo en cuatro fases principales, primero se configuró un objetivo específico para el estudio eligiendo las áreas críticas de la infraestructura y el personal objetivo, además se logró en primera instancia el compromiso

formal de la alta dirección de la empresa lo que permitió facilitar recursos técnicos y humanos se estructuró un equipo multidisciplinario con relación a las áreas de seguridad informática, redes, ética digital e inteligencia artificial.

En la fase de estimación inicial se realizó un ataque de simulación de ingeniería social y de ransomware para diagnosticar el nivel de amenaza de seguridad de la empresa. Posteriormente, se documentó y se seleccionó ChatGPT como herramienta, todo esto teniendo en cuenta sus capacidades técnicas, aplicaciones y sus límites en ética y normatividad. También se identificaron los requerimientos considerando infraestructura, software y conectividad para la implementación de sus usos de forma segura.

Durante la fase de experimentación, se diseñaron y ejecutaron dos tipos de campañas. En primer lugar, se ejecutó una campaña de phishing tradicional con IA no asistida y se registraron las tasas de éxito, las vulnerabilidades detectadas y los tiempos de respuesta, que fueron métricas de la organización. Luego se diseñó una segunda campaña utilizando mensajes generados por ChatGPT que replicaban los mismos escenarios, pero con un mayor grado de personalización e inteligencia contextual además se simularon escenarios específicos para ataques de ransomware emulando descargas de archivos captura de credenciales y acceso no autorizado sin comprometer realmente el entorno real.

Finalmente en la fase de análisis se realizó una evaluación comparativa desde una perspectiva cuantitativa y cualitativa se midieron indicadores como tasa de éxito de los ataques simulados tiempo de reacción del personal precisión de los mensajes y nivel de riesgo percibido. A nivel cualitativo, se analizó la percepción del personal frente a los ataques la complejidad de los mensajes generados y la capacidad de respuesta organizacional. Se documentaron también los riesgos técnicos, operativos, éticos y legales derivados del uso de IA generativa y se establecieron recomendaciones específicas para mitigar dichos riesgos garantizar el cumplimiento normativo y fortalecer la cultura de ciberseguridad organizacional.

Resultados

El impacto y alcance de esta investigación se centra en el uso tanto tradicional como asistido de IA para realizar ingeniería social en el contexto de penetración empresarial lo que se busca es medir el impacto y efectividad de los dos enfoques asociados en la exposición de las vulnerabilidades humanas en un contexto corporativo.

Las pruebas de penetración social se complementan cuando se aplican modelos de IA como los de ChatGPT. Tales herramientas aumentan de manera exponencial la efectividad de las campañas de penetración, phishing y fraude aumentando los niveles de interacción tipo de interacción por parte de los usuarios.

Los siguientes datos cuantitativos y cualitativos han sido organizados para facilitar su análisis comparativo y fundamentan las recomendaciones planteadas en las secciones posteriores del estudio.

Análisis de los Resultados

La IA más impactante fue ChatGPT desarrollada por OpenAI que dio la posibilidad de crear mensajes de phishing más elaborados con una construcción gramatical fluida usando información contextual de forma orgánica dentro de la información organizacional, en el caso de las empresas, a diferencia del método tradicional donde los mensajes eran genéricos y fácilmente activados por los filtros anti spam los correos electrónicos eran de un alto grado de persuasión e incluso imitaban el estilo de comunicación interna de la empresa.

El uso de esta tecnología permitió también incorporar elementos personalizados, como nombres de usuarios reales, referencias a sistemas internos y archivos simulados, aumentando la efectividad de la campaña. Estas capacidades, si bien útiles para la simulación, evidencian también el potencial riesgo si estas herramientas son empleadas maliciosamente.

El estudio se aplicó en la empresa Casystem, localizada en Cuenca, la cual cuenta con una infraestructura tecnológica de 10 hosts. Las pruebas se realizaron sobre una muestra de 15 colaboradores, seleccionados por su nivel de interacción con sistemas corporativos y acceso

a recursos críticos. Cada participante recibió correos simulados como parte de las campañas controladas.

A continuación, se presenta una tabla comparativa de los resultados obtenidos:

Tabla 1

Comparativa de resultados entre phishing tradicional y phishing con IA generativa

Métrica evaluada	Phishing tradicional	Phishing con IA generativa (ChatGPT)
Correos enviados	15	15
Usuarios que abrieron el correo	8 (53.3 %)	13 (86.6 %)
Clics en enlaces maliciosos	4 (26.6 %)	11 (73.3 %)
Usuarios que entregaron credenciales	2 (13.3 %)	8 (53.3 %)
Tiempo promedio de reacción (minutos)	21 min	43 min
Usuarios que reportaron el intento	5 (33.3 %)	1 (6.6 %)

Fuente: fuente propia.

Adicionalmente, se calculó un Índice de Riesgo Humano (IRH) como medida cuantitativa que expresa el porcentaje de usuarios que entregaron sus credenciales frente al total de correos enviados durante las campañas simuladas.

Este análisis se realizó mediante el uso de la siguiente formula

$$IRH = \frac{Número\ de\ usuarios\ que\ entregaron\ credenciales}{Total\ de\ correos\ enviados} \times 100$$

En la campaña de phishing tradicional

$$IRH = \frac{2}{15} \times 100 = 13.3\%$$

En la campaña de phishing con IA Generativa

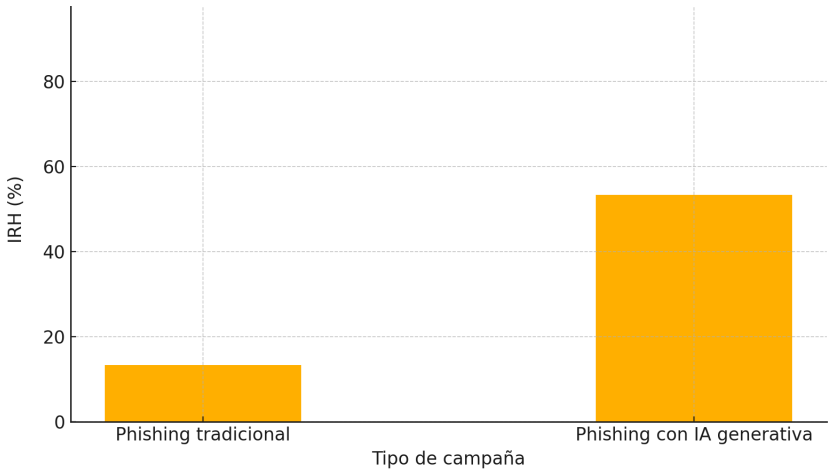
$$IRH = \frac{8}{15} \times 100 = 53.3\%$$

Este indicador proporciona una estimación clara del nivel de exposición humana ante técnicas de ingeniería social. En la campaña de phishing tradicional, el IRH fue de 13.3 %, al obtenerse 2 casos de entrega de credenciales sobre 15 intentos. En contraste, en la campaña asistida por ChatGPT, el IRH se elevó a 53.3 %, con 8 usuarios comprometidos de los mismos 15 correos enviados.

La información generada mediante inteligencia artificial no solo potencia la efectividad de los ataques simulados sino que también revela el creciente impacto que la inteligencia artificial puede desempeñar en el aumento de la susceptibilidad dentro de un entorno sumamente controlado, a diferencia observada entre ambos escenarios respalda el razonamiento de que la IA generativa puede suponer un considerable peligro en situaciones donde no se implemente una ética, normativa y estrategia definida.

Figura 1

Índice De Riesgo Humano (IRH) Por Tipo De Campaña



Fuente: fuente propia

El análisis de chi-cuadrado evidenció diferencias estadísticamente significativas entre las campañas.

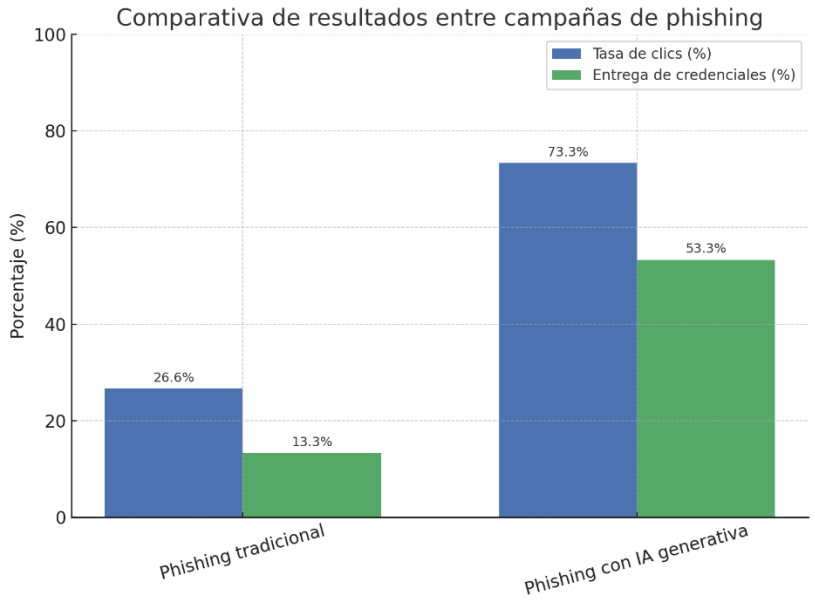
- Tasa de clics: campaña tradicional (26,6 %) vs. campaña con IA generativa (73,3 %), $\chi^2 = 6{,}48$, gl = 1, p = 0,011.
- Entrega de credenciales: campaña tradicional (13,3 %) vs. campaña con IA generativa (53,3 %), $\chi^2 = 4{,}84$, gl = 1, p = 0,028.

  Estos resultados confirman que el uso de IA generativa incrementa de forma notable la efectividad de los ataques simulados.

La Figura 2 muestra la comparación porcentual de la tasa de clics y la entrega de credenciales entre ambos enfoques. La Figura 3 representa el Índice de Riesgo Humano (IRH), evidenciando un aumento significativo en el escenario con IA generativa.

Figura 2

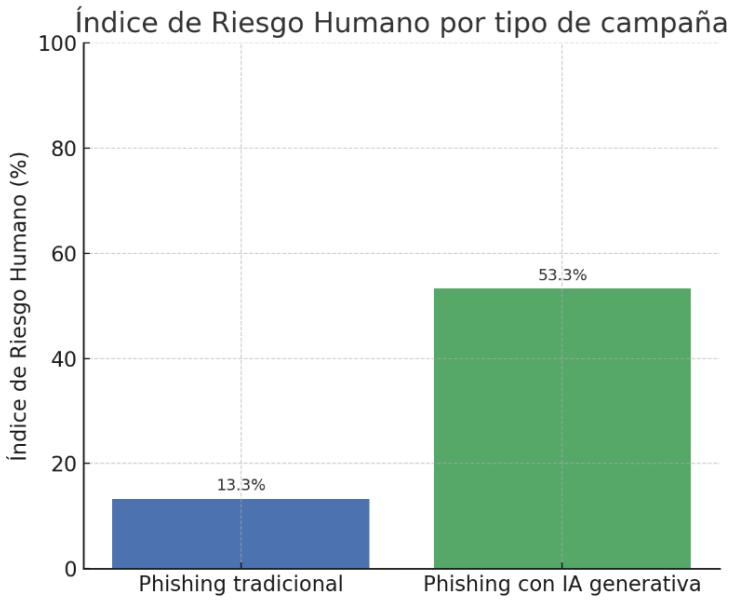
Comparativa de resultados entre campañas de phishing



Fuente: fuente propia.

Figura 3

Índice de riesgo Humano por tipo de campaña



Fuente: fuente propia.

Discusión

Los hallazgos de esta investigación permiten evidenciar mediante pruebas concretas, que la aplicación de herramientas de Inteligencia Artificial generativa en este caso ChatGPT modifica de manera muy efectiva las campañas de ingeniería social que se emplean en las pruebas de penetración, el principio más relevante que se ha logrado identificar es que la eficiencia que tiene la IA en la creación de mensajes simpáticos y contextualizados de muy alta persuasión aumenta de manera importante la exposición a ataques simulados en comparación con las pruebas manuales las cuales se ven limitadas por la IA. La IA incrementó los resultados de la ingeniería social mejorada por inteligencia artificial, con una tasa de entrega de credenciales del 53.3%, en comparación con el 13.3% para los métodos tradicionales. Además, se logró una tasa de clics del 73.3%, más del doble de lo que se observó en escenarios sin IA. Además, se notaron tiempos de reacción prolongados y tasas

de reporte más bajas en las campañas impulsadas por IA lo que implica que los encuestados eran no solo más vulnerables, sino menos conscientes del engaño.

Estos resultados coinciden con lo planteado por (Pasala, 2024), quien sostiene que los modelos de lenguaje generativo aumentan significativamente la efectividad de las campañas de phishing en entornos simulados. De la misma manera la investigación de Martínez (2024) demuestra que las pruebas de Pentesting asistidas por IA permiten identificar vulnerabilidades humanas con mayor precisión. Sin embargo en el presente estudio se alcanzaron cifras superiores en cuanto a compromiso de usuarios posiblemente debido a la personalización contextual del mensaje en función del entorno real de la empresa.

Por otra parte, los hallazgos también validan las advertencias realizadas por (Cordova et al., 2024), quien señaló que la IA generativa no solo representa una herramienta de mejora en ciberseguridad, sino también un riesgo emergente cuando es utilizada sin regulación o supervisión. En línea con esto se documentaron riesgos éticos relacionados con el consentimiento la manipulación psicológica y el uso de datos internos, incluso cuando se aplicaron de forma controlada.

Los hallazgos de la investigación sugieren que las organizaciones deben incluir en sus planes de ciberseguridad estrategias de formación más sofisticadas que tengan en cuenta el posible uso de la IA en ataques reales. Las campañas de concienciación deben evolucionar hacia simulaciones más complejas basadas en escenarios creados por IA con el propósito de poner a los empleados un paso adelante ante las amenazas del futuro.

El efecto de la campaña de concientización no fue medido, lo que podría ser una línea de investigación futura. Otra limitación es la pequeña muestra, que fue de solo 15 colaboradores en una sola organización, lo cual restringe la posibilidad de generalizar los resultados. Por más contundentes que sean estos hallazgos podrían ser totalmente distintos en las organizaciones que tengan diferentes culturas de seguridad políticas internas o niveles de madurez tecnológica.

Los resultados evidencian la falta de estrategias adecuadas en la capacitación en ciberseguridad a la par que se define el uso regulado de la IA en el ámbito corporativo. Esto, más la falta de normas claras sobre el uso ético de esta tecnología, requiere una profunda reconsideración. El uso no controlado de la IA, además, compromete su defensa. Por último, este trabajo concluye que, en el contexto del pentesting, la IA generativa no solo es efectiva, sino que logra poner en jaque los mecanismos de defensa que el ser humano ha históricamente construido

Conclusiones

Esta investigación demostró cómo la inteligencia artificial generativa aplicada a pruebas de penetración a través de técnicas de ingeniería social puede alterar asombrosamente los resultados obtenidos de tales ejercicios, se mostró a través de un análisis comparativo directo de campañas de phishing tradicionales a aquellas asistidas por IA como ChatGPT que el uso de inteligencia artificial aumentó significativamente la tasa de éxito en los ataques simulados del 13.3% al 53.3% en términos de envío de credenciales, así como también incrementó la tasa de clics en los enlaces maliciosos.

Con respecto al primer objetivo específico, se caracterizó con éxito el funcionamiento preciso de las herramientas generativas de IA, la capacidad de la inteligencia artificial para redactar mensajes específicos las hace flexibles y adaptadas al receptor. Esta habilidad resultó ser vital para el éxito de la campaña ya que los mensajes a los usuarios fueron identificados como reales por un mayor número de usuarios.

En base al segundo objetivo se observaron los beneficios técnico, operativo y estratégico. Con referencia a lo técnico la inteligencia artificial facilitó la generación automatizada y ágil de contenido de ataque. Desde el lado operativo la ejecución de las campañas se automatizó y se implementó con un alto grado de realismo, desde la perspectiva estratégica, se comprobó su utilidad como herramienta de diagnóstico para la detección de vulnerabilidades humanas que no se evidencian en pruebas tradicionales.

Respecto al tercer objetivo el caso práctico realizado en la empresa Casystem mostró que las exposiciones de vulnerabilidades a través de pruebas supervisadas por IA son superiores para las que están relacionadas con la acción y comportamiento humano la empresa tuvo baja capacidad de respuesta a los mensajes emitidos por inteligencia artificial lo que sugiere la necesidad urgente de reforzar la capacitación y la mejora a los sistemas de detección de alerta temprana.

Por último, en cuanto al cuarto objetivo se señalaron riesgos éticos y legales que deben ser tratados priorizados antes de la utilización de este tipo de herramientas en entornos laborales. La inteligencia artificial generativa en sus diferentes usos puede ser un apoyo en la mejora de la ciberseguridad, pero, al mismo, tiempo puede ser un peligro en caso que no exista un control de su uso, la realización de ataques incluso en entornos controlado debe ser enmarcada por el consentimiento informado, la proporcionalidad, la protección de datos, la supervisión institucional y otros similares conforme a lo que consta en la ley Orgánica de protección de datos personales (LOPDP) de Ecuador.

Para obtener el resultado final, esta investigación muestra tanto el inmenso cambio en el potencial de ataques de seguridad que la inteligencia artificial generativa trae a nuevas vulnerabilidades para la ciberdefensa (mayor riesgo de uso indebido). Por lo tanto, es uno de los mejores para ser utilizado en pruebas de penetración y solo en estándares específicos de la empresa que lo utiliza, entregando su gestión de datos a la compañía.

Trabajos futuros

Varias líneas de trabajo para estudios de seguimiento se derivan de los hallazgos de esta investigación. Incluyen, en primer lugar, aumentar la representatividad de los resultados. También debería ampliarse el tamaño y la diversidad de la muestra a otras áreas, niveles jerárquicos y organizaciones. En segundo lugar, sería valioso determinar si las campañas de promoción de la concienciación aumentaron el conocimiento de la recopilación y redujeron las respuestas dirigidas a los intentos de phishing de IA con el tiempo.

También se sugiere añadir otras medidas, como la capacidad de detección temprana de amenazas (utilizando inteligencia artificial defensiva), métricas de productividad y costos de implementación de este tipo de pruebas. También podríamos investigar qué tan bien funcionan diferentes modelos de IA generativa (GPT-4, LLaMA, Claude) y si hay diferencias significativas en las campañas.

Finalmente, se recomienda investigar el uso de marcos regulatorios internacionales (ISO/IEC 27001, NIST AI RMF) en la estructuración de pruebas de pentesting para el desarrollo de IA, con el objetivo de definir patrones convencionales y éticos que puedan ser reaplicados en diferentes entornos industriales.

Referencias bibliográficas

Cordova, R., Andrade-López, M., & Álvarez-Vera, M. (2024). Inteligencia artificial generativa en el ámbito de la ciberseguridad: una revisión sistemática de literatura. *MQRInvestigar*, *8*, 556–578. https://doi.org/10.56048/MQR20225.8.3.2024.556-578

EcuCERT. (2023). *Estadísticas de Seguridad de Redes de Telecomunicaciones*.

ENISA. (2022). *ENISA Annual Report on Cybersecurity Research and Innovation Needs and Priorities*. https://doi.org/10.13140/RG.2.2.11312.84484

Fitriana, D., Mas'udia, P., & Kusumawardani, M. (2023). NIST SP 800-115 Framework Implementation using Black Box Method on Security Gaps Testing on JTD Polinema's Official Website. *Jartel*, *13*, 328–335. https://doi.org/10.33795/jartel.v13i4.557

García, V., Zapata, E., Gómez, J., Laverde, L., & Macias, J. (2025). Modelo de gestión para la atención y respuesta ante ataques de ransomware en el área de networking. *Revista Sapientía*, *17*. https://doi.org/10.54278/sapientia.v17i33.263

Linares, F., & Salazar, B. (2025). Potencial y aplicaciones de la inteligencia artificial en seguridad. *Razón Crítica*, 1–16. https://doi.org/10.21789/25007807.2111

López, H., Zatarain, J., Solís, S., & Rendón, M. (2023). PERCEPCIÓN DE CIBERSEGURIDAD EN SISTEMAS DE INTELIGENCIA ARTIFICIAL EN LA EDUCACIÓN SUPERIOR. *Revista Digital de Tecnologías Informáticas y Sistemas*, *7*, 115–122. https://doi.org/10.61530/redtis.vol7.n1.2023.154.115-122

Martinez, C. (2024a). *TECNOLOGÍAS DE INTELIGENCIA ARTIFICIAL (IA) APLICADAS A LA SEGURIDAD INFORMÁTICA*. https://doi.org/10.13140/RG.2.2.12352.71681

Martínez, J. (2024). Impacto de la Inteligencia Artificial generativa en la publicación científica. *Enfermería Nefrológica*, *27*, 187–188. https://doi.org/10.37551/S2254-28842024019

NCSC. (2021). *Towards a Common ECSC Roadmap: Success factors for the implementation of national cyber security competitions*. https://doi.org/10.2824/657311

Pasala, R. (2024). USING GENERATIVE AI FOR ADAPTIVE INTRUSION DETECTION SYSTEMS. *International Journal of Engineering & Technology*, *Volume-08*, 7.

Perez, F. (2023). La inteligencia artificial generativa y su impacto en la creación de contenidos mediáticos. *Methaodos.Revista de Ciencias Sociales*, *11*, m231102a10. https://doi.org/10.17502/mrcs.v11i2.710

Sacoto, J., Machangara, O., & Araujo, J. (2025). El papel de la inteligencia artificial en la ciberseguridad de redes empresariales. *Revista Retos Para La Investigación*, *4*, 65–82. https://doi.org/10.62465/rri.v4n1.2025.126

Scarfone, K., Souppaya, M., Cody, A., & Orebaugh, A. (2021). *NIST Special Publication 800-115, Technical Guide to Information Security Testing and Assessment*.

Verizon DBIR. (2024). *2021 Verizon Data Breach Investigations Report*.

Conflicto de intereses:
Los autores declaran que no existe conflicto de interés posible.
Financiamiento:
No existió asistencia financiera de partes externas al presente artículo.
Agradecimiento:
Agradecimiento a Casystem y al a Universidad Católica de Cuenca. por permitirnos realizar este trabajo de investigación.
Nota:
El articulo no es producto de una publicación anterior.