

UNIVERSIDAD CATÓLICA DE CUENCA



Maestría en Ciberseguridad

**Informe de Investigación previo a la obtención del título de Magíster en
Ciberseguridad**

Tema: Impacto de los LLMs en Ciberseguridad: análisis comparativo entre
RAG y *Fine-Tuning* para la detección de vulnerabilidades

Autor: Ing. Leandro Antonio Pazmiño Ortiz

Asesores: Ing. Pablo Cuenca Tapia, Mg

Ing. Juan Carlos Ortega Castro, Mg

Cuenca, 2025

Certificación de Asesores

Se certifica que:

El informe de investigación “*Impacto de los LLMs en Ciberseguridad: análisis comparativo entre RAG y Fine-Tuning para la detección de vulnerabilidades*”, de autoría del señor Ingeniero en Electrónica y Telecomunicaciones Leandro Antonio Pazmiño Ortiz, CC: 1717115495, ecuatoriano, previo a la obtención del Título de Cuarto Nivel o Posgrado correspondiente a Magíster en Ciberseguridad cumple con la caracterización y estructura (parte protocolaria y parte expositiva) y se sujeta a la normativa pertinente exigida por el Consejo de Educación Superior, CES y la Universidad Católica de Cuenca, en consecuencia se autoriza su presentación para los trámites pertinentes.

Santa Ana de los Cuatro Ríos de Cuenca

Febrero, 2025.

Ing. Pablo Cuenca Tapia. Mg
Asesor Científico

Ing. Juan Carlos Ortega Castro. Mg
Asesor Metodológico

Certificación de Autoría

Certifico que:

“Impacto de los LLMs en Ciberseguridad: análisis comparativo entre RAG y *Fine-Tuning* para la detección de vulnerabilidades”, es el tema del informe final de investigación de mi AUTORÍA, previo a la obtención del Título de Cuarto Nivel o Posgrado correspondiente a Magíster en tecnologías de la información, por lo que, asumo su originalidad y el uso de fuentes de terceros registrados según las normas APA vigentes.

Santa Ana de los Cuatro Ríos de Cuenca

Febrero, 2025.

Ing. Leandro Antonio Pazmiño Ortiz. MSc.

CC: 171711545

Agradecimiento

A Dios, por brindarme la fuerza y la claridad necesarias para afrontar cada reto de esta etapa académica. Sin su guía y su bendición, este logro no habría sido posible.

A mis profesores cuyo conocimiento, y orientación fueron fundamentales en la construcción de este trabajo. A mis colegas y compañeros de estudio, por el enriquecedor intercambio de ideas y el apoyo mutuo en este proceso.

Finalmente, a mi familia, por ser mi pilar inquebrantable, por su amor incondicional y por creer en mí en todo momento. Su respaldo ha sido mi mayor motivación para seguir adelante.

Dedicatoria

A mis seres queridos, quienes con su amor, apoyo y comprensión han sido mi mayor motivación. A mis padres, por inculcarme valores de esfuerzo y dedicación; a mi familia, por su paciencia y aliento constante; y a mis amigos, por acompañarme en este viaje académico.

Dedico este trabajo a todos aquellos que, con su ejemplo y enseñanzas, me han inspirado a seguir creciendo en el conocimiento y en la vida.

Resumen

El creciente panorama de amenazas cibernéticas requiere estrategias avanzadas para detectar vulnerabilidades y salvaguardar infraestructuras críticas, revela una brecha en la aplicación sistemática de modelos de lenguaje de gran escala en dicho ámbito. La falta de estudios comparativos sobre la efectividad de Retrieval-Augmented Generation (RAG) y el ajuste fino (Fine-Tuning) acentúa la necesidad de análisis rigurosos que orienten su implementación. Esta investigación tuvo como objetivo analizar el desempeño de RAG y Fine-Tuning en la detección de vulnerabilidades para determinar su eficacia en distintos contextos de ciberseguridad. Se llevó a cabo un estudio de tipo bibliográfico que incluyó 35 artículos seleccionados mediante criterios de inclusión y exclusión claramente definidos. La muestra se basó en fuentes académicas indexadas, y el principal instrumento de recopilación fue una estrategia de búsqueda sistemática aplicada en bases de datos especializadas. El proceso contempló la clasificación de estudios por metodología, métricas de evaluación y aplicaciones prácticas, mientras que el análisis se efectuó comparando tasas de precisión, falsos positivos y capacidad de adaptación de cada técnica. Los resultados evidenciaron que ambas metodologías superaron a los métodos tradicionales en precisión (entre 80% y 90%) y redujeron significativamente la tasa de falsos positivos. RAG mostró mayor eficiencia en entornos dinámicos con acceso a datos externos, mientras que Fine-Tuning alcanzó un mejor rendimiento en dominios específicos con conjuntos de datos etiquetados. La adopción de RAG y Fine-Tuning aportan beneficios considerables en la identificación de amenazas, y su aplicación combinada posibilita abordar escenarios complejos con alta confiabilidad, contribuyendo así al fortalecimiento de la ciberseguridad.

Palabras claves: ciberseguridad, detección de vulnerabilidades, modelos de lenguaje de gran escala, *Retrieval-Augmented Generation*, *Fine-Tuning*.

Abstract

The expanding landscape of cyber threats requires advanced strategies for detecting vulnerabilities and safeguarding critical infrastructures, which reveals a gap in the systematic application of large-scale language models in this domain. The lack of comparative studies on the effectiveness of Retrieval-Augmented Generation (RAG) and Fine-Tuning underscores the need for rigorous analyses to guide their implementation. This study aimed to analyze the performance of RAG and Fine-Tuning in vulnerability detection in order to determine their effectiveness across diverse cybersecurity contexts. A bibliographic study was conducted, drawing on 35 articles selected through clearly defined inclusion and exclusion criteria. The sample was derived from indexed academic sources, and the primary instrument of data collection was a systematic search strategy applied in specialized databases. The process involved classifying studies by methodology, evaluation metrics, and practical applications, while the analysis focused on comparing precision rates, false positives, and each technique's adaptability. The results revealed that both methodologies outperformed traditional approaches in terms of precision (ranging from 80% to 90%) and significantly reduced the rate of false positives. RAG demonstrated higher efficiency in dynamic environments with access to external data, whereas Fine-Tuning yielded better performance in specific domains with labeled datasets. The adoption of RAG and Fine-Tuning provides considerable benefits for threat identification, and their combined application facilitates handling complex scenarios with high reliability, thereby contributing to the strengthening of cybersecurity.

Keywords: cybersecurity, vulnerability detection, large language models, *Retrieval-Augmented Generation*, *Fine-Tuning*.

Índice de contenidos

Capítulo I. Introducción	1
1.1 Situación problemática.....	1
1.2 Problema científico	2
1.3 Objeto de estudio	2
1.4 Campo de acción.....	3
1.5 Objetivos	3
1.6 Problema científico - preguntas de investigación	4
1.7 Variables	5
1.8 Justificación	6
1.9 Principales resultados a obtener.....	7
1.10 Marco teórico base	8
1.10.1 Relevancia de los Modelos de Lenguaje de Gran Escala (LLMs) en ciberseguridad ..	8
1.10.2 Modelos de Lenguaje de Gran Escala: transformación en ciberseguridad ...	9
1.10.3 Introducción a RAG y <i>Fine-Tuning</i>	9
1.11 Justificación de la investigación	10
1.12 Objetivos de la revisión bibliográfica	11
1.13 Marco teórico	11
1.13.1 Vulnerabilidades actuales en la seguridad cibernética.....	11
1.13.2 Modelos de Lenguaje de Gran Escala (LLMs) en ciberseguridad	13
1.13.3 Riesgos asociados con LLMs en ciberseguridad	14
1.13.4 Métodos avanzados de inteligencia artificial en ciberseguridad.....	15
1.13.5 <i>Retrieval-Augmented Generation</i> (RAG)	16
Definición y arquitectura de RAG	16
1.13.6 <i>Fine-Tuning</i> : proceso de ajuste y aplicaciones	17
Definición y proceso de <i>Fine-Tuning</i>	17
Capítulo II. Diagnóstico situacional.....	19
2.1 Antecedentes	19
2.1.1 Vulnerabilidades actuales en la seguridad cibernética.....	19
2.1.2 Modelos de Lenguaje de Gran Escala (LLMs)	20
2.1.3 Inteligencia artificial en detección de vulnerabilidades.....	20
2.1.4 Evaluación y mitigación de amenazas cibernéticas	21
2.2 Diseño y planificación de la revisión.....	22

2.3	Objetivo de la revisión	22
2.4	Estrategia de búsqueda.....	23
2.4.1	Fuentes primarias: bases de datos	23
2.4.2	Palabras clave y combinaciones.....	24
2.4.3	Filtros aplicados	25
2.4.4	Resultados iniciales.....	26
2.4.5	Ejemplo de aplicación de filtros y estrategias.....	27
2.4.6	Selección y análisis de fuentes.....	28
2.4.7	Criterios de inclusión y exclusión.....	28
2.4.8	Proceso de selección	31
2.4.9	Variables de estudio definidas	33
2.4.10	Detalle sobre las metodologías empleadas en los estudios seleccionados ..	34
2.4.11	Métricas utilizadas en los estudios.....	36
Capítulo III. Propuesta		37
3.1	Discusión de hallazgos y comparaciones.....	37
3.1.1	Análisis temático y comparativo de los estudios	37
3.1.2	Precisión en la detección de vulnerabilidades	37
3.1.3	Eficiencia operativa	37
3.1.4	Adaptabilidad a nuevas amenazas	38
3.2	Aplicaciones de <i>Fine-Tuning</i> en ciberseguridad.....	38
3.2.1	Casos de uso documentados de Fine Tuning	39
3.2.2	Limitaciones del <i>Fine-Tuning</i>	40
3.3	Aplicaciones de RAG en ciberseguridad	40
3.4	Comparación ampliada entre RAG y <i>Fine-Tuning</i>	43
3.5	Tendencias futuras y desafíos técnicos	44
3.6	Comparación entre RAG, <i>Fine-Tuning</i> y Métodos Tradicionales.....	45
3.7	Principales hallazgos comparativos	46
3.8	Contextos de uso recomendados según hallazgos	46
3.8.1	Contextos recomendados para RAG	47
3.8.2	Contextos recomendados para <i>Fine-Tuning</i>	48
3.8.3	Contextos donde los métodos tradicionales siguen siendo útiles	50
3.9	Discusión de limitaciones, desafíos y propuestas futuras.....	51
3.9.1	Limitaciones identificadas	51
3.10	Desafíos técnicos	52

3.11	Propuestas futuras para integrar de RAG y <i>Fine-Tuning</i> en ciberseguridad	53
3.11.1	Sistemas híbridos adaptativos	53
3.11.2	Automatización y democratización del <i>Fine-Tuning</i>	54
3.11.3	Inteligencia Artificial Explicable (XAI)	55
3.11.4	Aprendizaje federado y privacidad	55
3.11.5	Convergencia tecnológica y estándares	56
3.12	Conclusión de la discusión.....	56
	Conclusiones	59
	Recomendaciones.....	62
	Referencias bibliográficas.....	64

Índice de figuras

Figura 1	29
Figura 2	31

Índice de tablas

Tabla 1.....	43
Tabla 2.....	45

Capítulo I. Introducción

1.1 Situación problemática

En la actualidad, la ciberseguridad constituye un pilar esencial para garantizar la integridad, confidencialidad y disponibilidad de sistemas y datos en un mundo digital interconectado. Con el crecimiento de tecnologías emergentes como la inteligencia artificial, el Internet de las cosas (IoT) y la computación en la nube, las organizaciones enfrentan un panorama de amenazas cibernéticas cada vez más sofisticado y diverso. Las vulnerabilidades en infraestructuras críticas, datos sensibles y servicios esenciales exponen a empresas, gobiernos y usuarios individuales a riesgos significativos, como ataques de ransomware, campañas de phishing y la explotación de vulnerabilidades de día cero (Bhatt et al., 2024). La detección temprana y precisa de estas vulnerabilidades se ha convertido en una prioridad estratégica, especialmente porque los métodos tradicionales basados en firmas y reglas predefinidas han demostrado ser insuficientes para abordar las crecientes demandas de un entorno cibernético en constante cambio (Xu et al., 2024).

La evolución de las amenazas cibernéticas, caracterizada por ataques más frecuentes, específicos y técnicamente avanzados, resalta la necesidad de desarrollar herramientas y enfoques más eficaces para la detección y mitigación de riesgos. Los Modelos de Lenguaje de Gran Escala (LLMs, por sus siglas en inglés), como *Retrieval-Augmented Generation* (RAG) y el *Fine-Tuning*, emergen como soluciones prometedoras al combinar capacidades avanzadas de análisis con adaptabilidad a entornos cambiantes. Estas técnicas no solo incrementan la precisión en la identificación de vulnerabilidades, sino que también reducen significativamente los falsos positivos y mejoran los tiempos de respuesta ante incidentes críticos (Gennari et al., n.d.; Motlagh et al., 2024). Sin embargo, estas tecnologías enfrentan desafíos importantes, incluyendo

la necesidad de evaluar su efectividad en entornos reales y compararlas con métodos tradicionales para determinar las mejores prácticas y guías de implementación. Este escenario subraya la urgencia de llevar a cabo investigaciones exhaustivas y comparativas que impulsen el uso óptimo de estas herramientas avanzadas en la ciberseguridad moderna.

1.2 Problema científico

El problema científico que aborda este proyecto es la falta de evidencia comprensiva y comparativa sobre la efectividad de los modelos de lenguaje de gran escala (LLMs), específicamente *Retrieval-Augmented Generation* (RAG) y el *Fine-Tuning*, en la detección de vulnerabilidades en sistemas de ciberseguridad. A pesar del potencial teórico de estas tecnologías para superar las limitaciones de los métodos tradicionales, existe una carencia de estudios que evalúen sistemáticamente su desempeño en contextos prácticos y reales. La pregunta central que guía esta investigación es: ¿De qué manera los Modelos de Lenguaje de Gran Escala (LLMs) ajustados (*Fine-Tuning*) y basados en recuperación de información (RAG) pueden complementar los enfoques tradicionales de análisis de vulnerabilidades en la ciberseguridad?

Esta investigación busca llenar este vacío proporcionando una evaluación detallada y comparativa, basada en análisis cualitativos de estudios de caso y revisiones de la literatura existente, para así avanzar en el conocimiento y la práctica de la ciberseguridad.

1.3 Objeto de estudio

El objeto de estudio de este proyecto es el uso de modelos de lenguaje de gran escala (LLMs) en la detección de vulnerabilidades en sistemas de ciberseguridad. En particular, se enfocará en dos técnicas específicas: *Retrieval-Augmented Generation* (RAG) y el *Fine-Tuning* de LLMs. Estas técnicas representan enfoques avanzados en la aplicación de inteligencia artificial para identificar y mitigar vulnerabilidades, y su estudio pretende evaluar su efectividad

y practicidad en comparación con los métodos tradicionales utilizados en la industria de la ciberseguridad.

1.4 Campo de acción

El campo de acción de esta investigación se circunscribe al ámbito de la ciberseguridad, con un enfoque específico en la detección de vulnerabilidades utilizando modelos de lenguaje de gran escala (LLMs). Esto incluye la revisión y análisis de estudios académicos, reportes técnicos y estudios de caso que documentan la aplicación de técnicas como *Retrieval-Augmented Generation* (RAG) y el *Fine-Tuning* de LLMs en entornos reales de ciberseguridad. Además, se evaluarán los resultados y las metodologías de implementación de estas tecnologías para identificar sus ventajas, limitaciones y mejores prácticas. Este campo de acción permitirá un análisis profundo y contextualizado que pueda informar y guiar futuras investigaciones y aplicaciones prácticas en la detección de vulnerabilidades, contribuyendo al desarrollo de estrategias más efectivas y adaptables en la protección de sistemas informáticos.

1.5 Objetivos

General

Evaluar métodos avanzados de inteligencia artificial, como RAG y *Fine-Tuning* de LLMs, recopilando y revisando investigaciones académicas, estudios de caso y reportes técnicos relevantes para mejorar la detección de vulnerabilidades en sistemas de ciberseguridad.

Específicos

- Evaluar la efectividad de *Retrieval-Augmented Generation* (RAG) en la identificación de vulnerabilidades comparada con métodos tradicionales, mediante un análisis comparativo cuantitativo y cualitativo para determinar su

capacidad de recuperar y generar información relevante que identifique fallos de seguridad en sistemas de ciberseguridad.

- Investigar cómo el *Fine-Tuning* de Modelos de Lenguaje de Gran Escala (LLMs) afecta la precisión y eficiencia en la detección de vulnerabilidades específicas en ciberseguridad, con el fin de entender cómo la adaptación de estos modelos puede mejorar la capacidad de identificación de vulnerabilidades en sistemas críticos.
- Identificar las ventajas y desventajas de implementar RAG y *Fine-Tuning* de LLMs en entornos reales de ciberseguridad, con el propósito de proporcionar recomendaciones prácticas para la implementación efectiva de estas técnicas, basadas en estudios de caso y análisis cualitativos de su adaptabilidad y efectividad en diferentes contextos operativos.

1.6 Problema científico - preguntas de investigación

Preguntas científicas

Pregunta general de investigación

- ¿Cómo pueden los Modelos de Lenguaje de Gran Escala mejorar la detección de vulnerabilidades en sistemas de ciberseguridad?

Preguntas específicas de investigación

- ¿Cuál es la efectividad de *Retrieval-Augmented Generation* (RAG) en la identificación de vulnerabilidades comparada con métodos tradicionales?
- ¿Cómo afecta el *Fine-Tuning* de Modelos de Lenguaje de Gran Escala (LLMs) a la precisión y eficiencia en la detección de vulnerabilidades específicas en ciberseguridad?

- ¿Cuáles son las ventajas y desventajas de implementar RAG y *Fine-Tuning* de LLMs en entornos reales de ciberseguridad?

Hipótesis

General

Los modelos de lenguaje de gran escala, aplicados a través de técnicas como RAG y *Fine-Tuning*, mejorarán la detección de vulnerabilidades en sistemas de ciberseguridad en comparación con métodos tradicionales.

Hipótesis específicas

La técnica de *Retrieval-Augmented Generation* (RAG) identificará vulnerabilidades con mayor precisión y menor número de falsos positivos que los métodos tradicionales.

El *Fine-Tuning* de modelos de lenguaje de gran escala adaptados a conjuntos de datos específicos de ciberseguridad aumentará la eficiencia en la detección de amenazas dirigidas.

La implementación de RAG y *Fine-Tuning* en entornos reales de ciberseguridad demostrará ventajas significativas en términos de escalabilidad y adaptabilidad frente a métodos estándar de detección de vulnerabilidades.

1.7 Variables

Independiente:

- Técnica de detección de vulnerabilidades (RAG, *Fine-Tuning* de LLMs, Métodos Tradicionales).
- *Fine-Tuning* de Modelos de Lenguaje de Gran Escala (LLMs)
- RAG y *Fine-Tuning* de LLMs

Dependientes:

- Precisión en la detección de vulnerabilidades.
- Eficiencia en la detección de vulnerabilidades.
- Ventajas y desventajas en entornos reales de ciberseguridad.

1.8 Justificación

En el contexto actual de ciberseguridad, las amenazas evolucionan rápidamente, con ataques como el ransomware y los de día cero representando un desafío significativo para la protección de sistemas y datos sensibles (Bhatt et al., 2024). La detección temprana de vulnerabilidades es crucial, especialmente en sectores críticos como energía, salud y servicios financieros (Kairouz et al., 2021). Este estudio se enfoca en evaluar tecnologías avanzadas de inteligencia artificial para la detección de vulnerabilidades, específicamente los Modelos de Lenguaje de Gran Escala (LLMs).

Se investigan dos métodos avanzados: *Retrieval-Augmented Generation* (RAG) y ajuste fino (*Fine-Tuning*) de LLMs. RAG permite una respuesta más eficiente a incidentes cibernéticos integrando información relevante, mientras que el ajuste fino adapta los modelos a entornos específicos, mejorando la precisión y reduciendo falsos positivos (Ali, 2024; Abdullahi et al., 2022). La propuesta analiza comparativamente RAG y *Fine-Tuning* en entornos reales, evaluando sus ventajas y desventajas. La investigación proporcionará una comprensión profunda de la efectividad y aplicabilidad práctica de estas técnicas avanzadas, ofreciendo beneficios significativos en la mejora de la detección de vulnerabilidades y la protección de datos críticos (Mulder et al., n.d.; Gennari et al., n.d.; Motlagh et al., 2024).

1.9 Principales resultados a obtener

La investigación sobre la aplicación de Modelos de Lenguaje de Gran Escala (LLMs) en la detección de vulnerabilidades, centrada en la comparación entre *Retrieval-Augmented Generation* (RAG) y el ajuste fino (*Fine-Tuning*), tiene varios resultados clave esperados. Primero, se anticipa una evaluación detallada de la eficacia de ambos métodos en términos de precisión, tasa de detección de amenazas emergentes y reducción de falsos positivos. Este análisis comparativo proporcionará una visión clara de cuál de los dos métodos ofrece un mejor desempeño en diferentes contextos cibernéticos, facilitando la selección de la técnica más adecuada para diversas aplicaciones.

En segundo lugar, se espera un examen exhaustivo del impacto de cada técnica en la capacidad de respuesta ante incidentes cibernéticos. Se evaluará cómo RAG contribuye a la integración de información relevante y la generación de respuestas precisas, mientras que el ajuste fino se analizará en términos de su capacidad para personalizar modelos adaptados a entornos específicos. Este análisis permitirá optimizar la respuesta a incidentes y mejorar la gestión de amenazas.

La investigación también ofrecerá un análisis detallado de la aplicabilidad de estas técnicas en sectores críticos como energía, salud y servicios financieros. Se proporcionarán recomendaciones basadas en la efectividad de RAG y ajuste fino en estos contextos, orientadas a mejorar la protección de infraestructuras esenciales.

Finalmente, se espera que los resultados contribuyan a una comprensión más profunda de cómo estas técnicas avanzadas de LLMs pueden fortalecer la detección y mitigación de vulnerabilidades. Este conocimiento no solo mejorará las defensas cibernéticas y la protección de

datos críticos, sino que también ofrecerá una base sólida para futuras investigaciones y desarrollos en el campo de la ciberseguridad.

1.10 Marco teórico base

1.10.1 Relevancia de los Modelos de Lenguaje de Gran Escala (LLMs) en ciberseguridad

En la era digital, las organizaciones enfrentan un panorama cada vez más complejo de amenazas cibernéticas. Desde ataques de ransomware hasta vulnerabilidades de día cero, el aumento en la sofisticación y frecuencia de los ataques compromete la seguridad de infraestructuras críticas, datos sensibles y operaciones empresariales (Bhatt et al., 2024). Ante este desafío, la detección de vulnerabilidades se ha convertido en una tarea prioritaria dentro de la ciberseguridad, ya que constituye la primera línea de defensa contra posibles ataques (Gennari et al., n.d.). Sin embargo, los métodos tradicionales de detección, como los sistemas basados en firmas y reglas predefinidas, presentan limitaciones significativas en términos de precisión, adaptabilidad y escalabilidad, especialmente en un entorno donde las amenazas evolucionan rápidamente (Xu et al., 2024).

Los Modelos de Lenguaje de Gran Escala (LLMs, por sus siglas en inglés) han emergido como una solución innovadora en este contexto. Su capacidad para analizar patrones complejos en el lenguaje natural, procesar grandes volúmenes de datos y generar contenido coherente ha revolucionado campos como el procesamiento de lenguaje natural (PLN) y ahora se perfila como una herramienta poderosa en la ciberseguridad (Ali, 2024). Dentro de este marco, dos técnicas específicas destacan por su potencial: *Retrieval-Augmented Generation* (RAG) y el *Fine-Tuning* de LLMs. Ambas ofrecen enfoques complementarios para abordar las limitaciones de los

métodos tradicionales, mejorando la precisión, eficiencia y adaptabilidad en la detección de vulnerabilidades.

1.10.2 Modelos de Lenguaje de Gran Escala: transformación en ciberseguridad

Los LLMs, como GPT-3 y BERT, han demostrado ser herramientas altamente efectivas en tareas de PLN debido a su capacidad para comprender y generar lenguaje natural de manera contextual. En ciberseguridad, su aplicación se extiende a áreas como la generación de simulaciones de ataques, la identificación de amenazas emergentes y la mejora de sistemas de detección de intrusiones (Ali, 2024). Estas capacidades se deben a su arquitectura basada en redes neuronales profundas, que les permite analizar patrones complejos y correlacionar grandes cantidades de datos de manera eficiente.

Un estudio reciente realizado por Xu et al. (2024) destacó que los LLMs pueden reducir el tiempo necesario para identificar vulnerabilidades críticas en un 30%, en comparación con los métodos tradicionales. Además, su capacidad para generar reportes automatizados y realizar análisis contextuales en tiempo real los convierte en una solución integral para abordar las necesidades actuales de ciberseguridad.

1.10.3 Introducción a RAG y *Fine-Tuning*

Dentro del ecosistema de los LLMs, dos enfoques específicos se destacan por su aplicabilidad en la detección de vulnerabilidades:

Retrieval-Augmented Generation (RAG): Combina las capacidades generativas de los LLMs con sistemas de recuperación de información, lo que permite integrar datos externos en las respuestas generadas. En ciberseguridad, esta técnica puede utilizarse para correlacionar

incidentes con bases de datos de amenazas conocidas, como CVE (Common Vulnerabilities and Exposures), mejorando la precisión y reduciendo los falsos positivos (Motlagh et al., 2024).

Fine-Tuning: Ajusta modelos preentrenados utilizando conjuntos de datos específicos de un dominio, adaptándolos a las necesidades particulares de una organización o sector. Por ejemplo, en infraestructuras críticas como salud o energía, el *Fine-Tuning* puede personalizar la detección de vulnerabilidades para enfrentar amenazas específicas (Bhatt et al., 2024).

Ambos enfoques ofrecen ventajas significativas en comparación con los métodos tradicionales, aunque también presentan desafíos que deben ser abordados.

1.11 Justificación de la investigación

La incorporación de técnicas avanzadas como RAG y *Fine-Tuning* en la ciberseguridad promete transformar la manera en que se detectan y mitigan las vulnerabilidades. Sin embargo, a pesar de su potencial, existen pocos estudios que evalúen de manera comprensiva y comparativa su efectividad en contextos reales (Gennari et al., n.d.). Esta falta de evidencia representa un vacío crítico en la literatura, especialmente cuando se considera la importancia de estas técnicas para enfrentar amenazas cada vez más sofisticadas.

Esta investigación busca llenar ese vacío mediante una revisión bibliográfica exhaustiva y un análisis comparativo cualitativo, centrándose en métricas clave como precisión, tasa de falsos positivos, eficiencia y escalabilidad. Los resultados no solo proporcionarán una comprensión más profunda de las capacidades de RAG y *Fine-Tuning*, sino que también ofrecerán guías prácticas para su implementación en la industria.

1.12 Objetivos de la revisión bibliográfica

El objetivo principal de esta revisión es establecer una base teórica sólida sobre RAG y *Fine-Tuning* en ciberseguridad, analizando sus aplicaciones, ventajas, limitaciones y efectividad en comparación con los métodos tradicionales. Específicamente, esta investigación tiene como objetivo identificar y sintetizar estudios relevantes sobre el uso de *Retrieval-Augmented Generation* (RAG) y el *Fine-Tuning* en la detección de vulnerabilidades en sistemas de ciberseguridad. Esto implica realizar una revisión exhaustiva de la literatura disponible para comprender cómo estas técnicas han sido aplicadas, evaluando sus fortalezas, limitaciones y el impacto que tienen en el campo.

Además, se busca comparar ambas técnicas utilizando criterios de evaluación específicos, como la precisión y la eficiencia. Este análisis comparativo permitirá determinar cuál de los enfoques resulta más efectivo en diferentes escenarios, proporcionando datos concretos que puedan servir como base para futuras implementaciones en la industria. Por último, se pretende ofrecer una visión general de las tendencias actuales y las mejores prácticas en el uso de Modelos de Lenguaje de Gran Escala (LLMs) en Ciberseguridad. Este apartado se centrará en destacar innovaciones recientes, casos de éxito y recomendaciones clave para optimizar el uso de estas herramientas en la protección de sistemas frente a amenazas emergentes.

1.13 Marco teórico

1.13.1 Vulnerabilidades actuales en la seguridad cibernética

Definición y tipología de vulnerabilidades

Las vulnerabilidades en la seguridad cibernética se refieren a debilidades o fallos en los sistemas informáticos, redes, aplicaciones y protocolos que pueden ser explotados por atacantes

para comprometer la integridad, confidencialidad y disponibilidad de los datos y servicios. Estas vulnerabilidades pueden clasificarse en varias categorías, como vulnerabilidades de software, errores de configuración, problemas de gestión de identidad y acceso, y vulnerabilidades físicas(*Sophos 2024 Threat Report: Cybercrime on Main Street*, n.d.). La clasificación de estas vulnerabilidades es crucial para el desarrollo de estrategias de mitigación adecuadas y la implementación de controles de seguridad eficaces.

Además, las vulnerabilidades de software suelen ser el resultado de errores de programación, como desbordamientos de búfer, inyecciones de código y fallos en la validación de entradas. Por otro lado, los errores de configuración incluyen configuraciones predeterminadas inseguras y permisos inapropiados. Los problemas de gestión de identidad y acceso pueden derivarse de contraseñas débiles, gestión inadecuada de privilegios y fallos en la autenticación. Las vulnerabilidades físicas, aunque menos comunes, abarcan el acceso no autorizado a hardware y la manipulación de dispositivos. Comprender y clasificar estas vulnerabilidades permite a las organizaciones priorizar sus esfuerzos de seguridad y proteger mejor sus activos críticos (*Sophos 2024 Threat Report: Cybercrime on Main Street*, n.d.).

Evolución de las amenazas y explotación de vulnerabilidades

La evolución de las amenazas cibernéticas ha llevado a un aumento en la sofisticación y frecuencia de los ataques, como ransomware, phishing, y ataques de día cero (*Sophos 2024 Threat Report: Cybercrime on Main Street*, n.d.). Las vulnerabilidades de día cero son particularmente peligrosas, ya que explotan fallos no conocidos previamente y para los cuales no existen parches disponibles. Estos ataques pueden causar interrupciones significativas en los servicios y resultar en la pérdida de datos sensibles, impactando negativamente a las

organizaciones y sus clientes. La detección y mitigación temprana de estas vulnerabilidades es esencial para reducir el impacto potencial de los ataques.

Además, la creciente interconexión y digitalización de infraestructuras críticas aumentan la superficie de ataque y el riesgo de explotación. Sectores como la energía, la salud y los servicios financieros se han convertido en objetivos atractivos para los atacantes debido a la dependencia de sistemas tecnológicos avanzados y la gran cantidad de datos sensibles que manejan. La colaboración entre gobiernos, empresas y organizaciones de seguridad es fundamental para desarrollar soluciones integrales que aborden estas amenazas emergentes. Las inversiones en investigación y desarrollo de tecnologías avanzadas de ciberseguridad son cruciales para mantener la resiliencia frente a la evolución constante de las amenazas cibernéticas.

1.13.2 Modelos de Lenguaje de Gran Escala (LLMs) en ciberseguridad

Aplicaciones de LLMs en detección de vulnerabilidades

LLMs como GPT-3 y BERT se utilizan para mejorar la detección de vulnerabilidades debido a su capacidad para analizar patrones complejos en el lenguaje natural (Xu et al., 2024a). Estas tecnologías permiten automatizar el análisis de código fuente y detectar anomalías que podrían indicar la presencia de vulnerabilidades de seguridad. Iniciativas como RepairLlama y Hackmentor demuestran la adaptabilidad de los LLMs para crear modelos específicos de ciberseguridad. Estas herramientas son capaces de escanear grandes volúmenes de código, identificar posibles vulnerabilidades y proponer soluciones de manera eficiente.

Además, los LLMs pueden ser entrenados con grandes conjuntos de datos de vulnerabilidades conocidas y patrones de ataques para mejorar su capacidad de detección. La

integración de LLMs en los procesos de desarrollo de software y pruebas de seguridad puede reducir significativamente el tiempo y el esfuerzo necesarios para identificar y corregir fallos de seguridad (Bhatt et al., 2024a). Las empresas de ciberseguridad están explorando continuamente nuevas formas de utilizar LLMs para aumentar la eficacia de sus soluciones de seguridad y proporcionar una defensa más robusta contra las amenazas cibernéticas emergentes.

1.13.3 Riesgos asociados con LLMs en ciberseguridad

El uso de LLMs en generación de código puede introducir vulnerabilidades no intencionadas, comprometiendo aplicaciones críticas (Mulder et al., n.d.). Los modelos generativos pueden producir código que, aunque funcional, contiene errores de seguridad debido a la falta de comprensión contextual o a la imitación de patrones inseguros presentes en los datos de entrenamiento. La manipulación de LLMs por actores maliciosos para generar deep fakes y desinformación plantea serios riesgos de seguridad (Mulder et al., n.d.). Estos modelos pueden ser explotados para crear contenido engañoso, manipular opiniones públicas y llevar a cabo ataques de ingeniería social.

Además, la falta de interpretabilidad y transparencia en los modelos puede dificultar la detección y mitigación de amenazas generadas (Abdullahi et al., 2022a). La naturaleza de caja negra de los LLMs hace que sea desafiante entender cómo y por qué toman ciertas decisiones, lo que complica la identificación de posibles fallos de seguridad y la implementación de correcciones adecuadas. Es crucial desarrollar enfoques que permitan una mayor transparencia y explicabilidad en los modelos de IA para asegurar su uso seguro y fiable en aplicaciones de ciberseguridad. La investigación en inteligencia artificial explicable y la implementación de prácticas de desarrollo ético son fundamentales para abordar estos desafíos.

1.13.4 Métodos avanzados de inteligencia artificial en ciberseguridad

Retrieval-Augmented Generation (RAG)

La combinación de RAG con LLMs ofrece nuevas perspectivas en la identificación y mitigación de vulnerabilidades (Motlagh et al., 2024a). RAG permite la integración eficiente de información dispersa y mejora la detección de amenazas emergentes (Sewak et al., 2023a). Esta metodología combina técnicas de recuperación de información con capacidades generativas de los LLMs para proporcionar respuestas contextualizadas y precisas en tiempo real. La capacidad de RAG para manejar grandes volúmenes de datos y extraer información relevante permite una detección más rápida y efectiva de amenazas.

Además, la implementación de RAG en sistemas de ciberseguridad puede mejorar significativamente la capacidad de respuesta ante incidentes. Al combinar la generación de contenido con la recuperación de datos específicos, RAG puede proporcionar a los analistas de seguridad información crítica de manera oportuna, facilitando la toma de decisiones informadas. Este enfoque integrado puede reducir el tiempo necesario para identificar y mitigar amenazas, mejorando la resiliencia de los sistemas frente a ataques cibernéticos. La investigación continua en RAG y su aplicación en ciberseguridad promete avances significativos en la protección de infraestructuras críticas y datos sensibles.

Sistemas de Detección De Intrusiones (IDS) de próxima generación

El uso de LLMs en IDS permite una detección de anomalías en tiempo real más precisa y adaptable (Ali, 2024a). Los IDS avanzados basados en LLMs integran inteligencia artificial explicable, aumentando la transparencia y confianza en los sistemas de seguridad (Ali, 2024a). La incorporación de técnicas de aprendizaje profundo y redes neuronales en IDS ha demostrado

mejorar significativamente la capacidad de detección y respuesta ante ciberataques. Estos sistemas pueden analizar patrones de tráfico de red y comportamientos de usuarios para identificar actividades sospechosas y potenciales amenazas de manera proactiva.

Además, los IDS de próxima generación son capaces de adaptarse a nuevas amenazas a medida que emergen, gracias a su capacidad de aprendizaje continuo. Esto permite que los sistemas se mantengan actualizados frente a técnicas de ataque evolucionadas y cambiantes. La integración de LLMs en IDS también facilita la generación de alertas contextuales y recomendaciones de mitigación específicas, lo que ayuda a los equipos de seguridad a responder de manera más efectiva a incidentes. La combinación de detección avanzada, aprendizaje adaptativo y explicabilidad hace que estos IDS sean una herramienta poderosa en la defensa contra ciberataques sofisticados.

1.13.5 *Retrieval-Augmented Generation (RAG)*

Definición y arquitectura de RAG

Retrieval-Augmented Generation (RAG) es una metodología híbrida innovadora que integra dos componentes fundamentales para mejorar la eficiencia y precisión de las respuestas generadas por sistemas de inteligencia artificial. Esta técnica se estructura de la siguiente manera:

1. **Sistema de recuperación (Retriever):** Este componente es esencial para el proceso, ya que se encarga de recuperar información pertinente de diversas fuentes externas. Estas fuentes pueden incluir bases de datos de amenazas, reportes de vulnerabilidades, y documentos técnicos, entre otros. La capacidad de acceder a una amplia gama de datos asegura que la información utilizada sea relevante y actual.

2. **Modelo generativo (Generator):** Una vez que la información ha sido recuperada, el componente generativo, que usualmente es un Modelo de Lenguaje de Gran Escala (LLM) preentrenado, toma esta información y genera respuestas contextualizadas. Este proceso no solo se basa en la información recuperada, sino que también adapta las respuestas para que sean coherentes y específicas al contexto en el que se utilizan.

La fusión de estos dos componentes confiere a RAG una capacidad única para generar respuestas que no solo son precisas, sino también extremadamente pertinentes y actualizadas, lo que es crucial en campos como la ciberseguridad. En este contexto, RAG permite una correlación efectiva entre eventos de seguridad y datos en tiempo real, mejorando significativamente la detección y respuesta a amenazas emergentes, como destacaron Lewis et al. en 2020. Esta capacidad de integrar y contextualizar información en tiempo real convierte a RAG en una herramienta poderosa para enfrentar los desafíos dinámicos y en constante evolución de la ciberseguridad moderna.

1.13.6 *Fine-Tuning*: proceso de ajuste y aplicaciones

Definición y proceso de *Fine-Tuning*

El *Fine-Tuning*, también conocido como ajuste fino, consiste en la adaptación de modelos de lenguaje preentrenados a tareas específicas a través del uso de datos altamente especializados. Este proceso es esencial en campos como la ciberseguridad, donde permite personalizar modelos avanzados como GPT-3 o BERT para funciones críticas como la detección de vulnerabilidades, el análisis de patrones de ataques y la generación de respuestas adaptadas a incidentes específicos.

El proceso de *Fine-Tuning* incluye varias etapas clave:

1. **Selección y preparación de datos:** Esta fase es fundamental y requiere la recopilación y el procesamiento de conjuntos de datos que sean representativos del dominio específico al que se aplicará el modelo. Esto asegura que el modelo se entrene con información relevante y contextual.
2. **Entrenamiento supervisado:** En esta etapa, el modelo preentrenado se ajusta utilizando los datos especializados preparados previamente. Esta adaptación se realiza a través de múltiples iteraciones, optimizando los parámetros del modelo para mejorar su rendimiento en la tarea específica.
3. **Validación y evaluación:** Finalmente, el desempeño del modelo ajustado se evalúa utilizando métricas específicas como la precisión, eficiencia y tasa de falsos positivos. Esta evaluación es crucial para verificar la efectividad del modelo en el entorno real y ajustar cualquier parámetro necesario antes de su despliegue final.

El *Fine-Tuning*, o ajuste fino, es un proceso que adapta modelos de lenguaje preentrenados para tareas específicas mediante el uso de datos especializados. En el ámbito de la ciberseguridad, el *Fine-Tuning* permite personalizar modelos como GPT-3 o BERT para detectar vulnerabilidades, analizar patrones de ataques y generar respuestas específicas para incidentes críticos.

Capítulo II. Diagnóstico situacional

2.1 Antecedentes

2.1.1 Vulnerabilidades actuales en la seguridad cibernética

La necesidad de proteger los sistemas contra vulnerabilidades explotables ha llevado a la adopción de métodos de inteligencia artificial para detectar y prevenir ciberataques (Abdullahi et al., 2022a). Las vulnerabilidades en áreas críticas como la salud y la energía destacan la importancia de la ciberseguridad y la necesidad de enfoques avanzados para la mitigación de riesgos (Abdullahi et al., 2022a). La evaluación continua y la gestión de parches son esenciales para minimizar los riesgos de seguridad. Las organizaciones deben implementar políticas de gestión de vulnerabilidades que incluyan la identificación, evaluación y remediación de fallos de seguridad de manera proactiva.

Además, el desarrollo de tecnologías y herramientas de escaneo de vulnerabilidades basadas en IA ha permitido una detección más precisa y rápida de fallos de seguridad. Estas herramientas utilizan algoritmos avanzados para analizar grandes volúmenes de datos y detectar patrones que indican la presencia de vulnerabilidades (Gennari et al., n.d.). La colaboración entre la academia, la industria y los organismos gubernamentales es fundamental para avanzar en la investigación y desarrollo de soluciones de ciberseguridad. La creación de marcos regulatorios y estándares de seguridad también juega un papel crucial en la protección de infraestructuras críticas y la reducción de riesgos.

2.1.2 Modelos de Lenguaje de Gran Escala (LLMs)

LLMs como GPT-3, BERT y otros han demostrado ser efectivos en tareas de comprensión y generación de lenguaje, mejorando la ciberseguridad (Xu et al., 2024a). Los LLMs se utilizan para crear escenarios de ejercicios de ciberseguridad y mejorar la preparación ante amenazas complejas (Yamin et al., 2024a). Estos modelos pueden generar simulaciones realistas de ataques cibernéticos, ayudando a los equipos de seguridad a desarrollar y practicar estrategias de respuesta efectivas. Además, su capacidad para analizar grandes volúmenes de datos y extraer información relevante los hace herramientas valiosas para la inteligencia de amenazas (Bhatt et al., 2024a).

Además, los LLMs pueden ser empleados para la creación de informes automatizados y la generación de alertas sobre actividades sospechosas en tiempo real. Su habilidad para procesar y comprender lenguaje natural permite la integración en sistemas de respuesta a incidentes, facilitando la identificación y mitigación de amenazas. La investigación en la mejora de la precisión y eficiencia de los LLMs continúa, con el objetivo de desarrollar modelos más robustos y adaptativos. La implementación de LLMs en ciberseguridad promete avances significativos en la protección de sistemas y datos sensibles frente a una amplia gama de amenazas cibernéticas.

2.1.3 Inteligencia artificial en detección de vulnerabilidades

Retrieval-Augmented Generation (RAG) y el ajuste fino de LLMs son métodos avanzados para detectar y mitigar vulnerabilidades (Motlagh et al., 2024a). CYBERSECEVAL 2 es una suite de evaluación que cuantifica los riesgos y capacidades de seguridad de los LLMs (Bhatt et al., 2024a). Estas técnicas permiten una respuesta proactiva y eficiente a las amenazas cibernéticas, mejorando la seguridad de los sistemas informáticos (Zhang et al., 2024a). La

capacidad de RAG para integrar información relevante y generar contenido específico permite a los analistas de seguridad responder de manera más efectiva a incidentes y mitigar riesgos de manera oportuna.

Además, el ajuste fino de LLMs permite personalizar los modelos para adaptarse a los entornos y necesidades específicas de las organizaciones. Esta personalización puede mejorar la precisión y relevancia de las detecciones, reduciendo los falsos positivos y mejorando la eficiencia operativa. La colaboración entre investigadores y profesionales de la ciberseguridad es esencial para desarrollar y perfeccionar estas técnicas avanzadas. La continua innovación en el uso de inteligencia artificial para la detección y mitigación de vulnerabilidades promete fortalecer significativamente las defensas cibernéticas y proteger mejor a las organizaciones contra amenazas emergentes.

2.1.4 Evaluación y mitigación de amenazas cibernéticas

Los modelos de lenguaje multimodal plantean tanto oportunidades como desafíos en ciberseguridad, desde la generación de contenido hasta la prevención de abusos tecnológicos (Mulder et al., n.d.). La evaluación de LLMs en tareas de ciberseguridad ayuda a identificar potenciales vulnerabilidades y mejorar la seguridad de los sistemas (Gennari et al., n.d.). Los modelos multimodales, que combinan texto, imágenes y otros datos, pueden proporcionar una visión más completa y precisa de las amenazas, mejorando la capacidad de detección y respuesta. Sin embargo, también es crucial abordar los riesgos asociados con el uso de estos modelos para garantizar su uso seguro y ético.

Además, la incorporación de técnicas de aprendizaje federado y privacidad diferencial en LLMs promete aumentar la seguridad y privacidad en las aplicaciones de ciberseguridad

(Kairouz et al., 2021). El aprendizaje federado permite el entrenamiento de modelos en datos distribuidos sin necesidad de centralizar la información, reduciendo el riesgo de exposición de datos sensibles. La privacidad diferencial proporciona garantías matemáticas sobre la privacidad de los datos, asegurando que la información individual no pueda ser inferida a partir de los resultados del modelo. Estas tecnologías emergentes ofrecen nuevas formas de mejorar la seguridad y privacidad de los sistemas de ciberseguridad, contribuyendo a la protección de datos críticos y la reducción de riesgos.

2.2 Diseño y planificación de la revisión

La planificación de una revisión bibliográfica implica establecer un enfoque estructurado para garantizar que los resultados sean relevantes, exhaustivos y representativos de los avances más recientes en el campo. Este proceso es esencial para proporcionar una base sólida que respalde la investigación sobre el uso de *Retrieval-Augmented Generation* (RAG) y *Fine-Tuning* en la ciberseguridad.

2.3 Objetivo de la revisión

El objetivo central de este estudio consiste en identificar, analizar y sintetizar las investigaciones que aborden el uso de *Retrieval-Augmented Generation* (RAG) y *Fine-Tuning* para la detección de vulnerabilidades en el ámbito de la ciberseguridad. Esta indagación se estructura alrededor de tres preguntas fundamentales:

1. ¿Cuáles son las implementaciones concretas de RAG y *Fine-Tuning* que se han llevado a cabo en el campo de la ciberseguridad? Este enfoque busca explorar la diversidad y el alcance de las aplicaciones prácticas de estas tecnologías avanzadas en la mejora de la seguridad informática.

2. ¿Cómo se comparan estas técnicas innovadoras con los métodos tradicionales en términos de precisión, eficiencia y adaptabilidad? El análisis se enfoca en evaluar el rendimiento relativo de RAG y *Fine-Tuning* frente a enfoques más establecidos, proporcionando una base comparativa que destaque ventajas y posibles mejoras.
3. ¿Cuáles son los principales desafíos y limitaciones que estas técnicas enfrentan según la literatura existente? Esta pregunta guía la identificación de las barreras comunes y los obstáculos técnicos que podrían afectar la adopción y la efectividad de RAG y *Fine-Tuning* en entornos reales de ciberseguridad.

Este marco de análisis no solo enriquece la comprensión de las tecnologías emergentes en ciberseguridad, sino que también orienta futuras investigaciones hacia áreas críticas de desarrollo y aplicación.

2.4 Estrategia de búsqueda

2.4.1 Fuentes primarias: bases de datos

Para garantizar la calidad y relevancia de las fuentes, se seleccionaron cinco bases de datos académicas ampliamente reconocidas por su prestigio y especialización. En primer lugar, IEEE Xplore fue elegida por su enfoque en tecnología e ingeniería, ofreciendo una amplia gama de publicaciones relacionadas con inteligencia artificial, que resulta esencial para temas avanzados como los *Large Language Models* (LLMs).

En segundo lugar, la ACM Digital Library se incluyó por su especialización en ciencias de la computación y su enfoque en aplicaciones prácticas de LLMs, lo que la convierte en una fuente clave para investigaciones técnicas. Por otro lado, Scopus destaca por su cobertura

multidisciplinaria, proporcionando acceso a artículos revisados por pares en diversos campos, lo cual asegura un panorama amplio y robusto para el análisis.

Asimismo, Google Scholar se utilizó por su capacidad para identificar estudios recientes, incluyendo preprints y documentos técnicos que ofrecen información actualizada y perspectivas emergentes en el área. Finalmente, se incorporó PubMed debido a su relevancia en temas de salud, especialmente en el uso de LLMs para la seguridad en sistemas críticos, un aspecto crucial en aplicaciones relacionadas con la medicina y la protección de datos sensibles.

2.4.2 Palabras clave y combinaciones

Para optimizar la búsqueda y asegurar la cobertura exhaustiva de la temática, se definieron combinaciones específicas de términos clave orientadas a explorar diversas facetas de la investigación:

- "*Retrieval-Augmented Generation* AND cybersecurity": Esta combinación de términos se centró en identificar estudios que abordan el uso de RAG en el contexto de la ciberseguridad, proporcionando una perspectiva sobre las aplicaciones específicas y los beneficios directos de esta tecnología.
- "*Fine-Tuning* AND large language models AND vulnerability detection": Esta frase de búsqueda se diseñó para localizar investigaciones que examinen cómo el ajuste fino de modelos de lenguaje grandes puede ser aplicado para detectar vulnerabilidades, una área de interés creciente dentro de la seguridad informática.
- "LLMs AND threat detection AND comparison with traditional methods": Se utilizó esta búsqueda para hallar documentos que comparan la detección de amenazas mediante

modelos de lenguaje de gran escala con métodos tradicionales, buscando evaluar la eficiencia y efectividad relativas.

- "Cybersecurity applications of AI AND RAG/*Fine-Tuning*": Finalmente, esta combinación busca trabajos que discutan aplicaciones de inteligencia artificial, específicamente RAG y *Fine-Tuning*, dentro de la ciberseguridad, abarcando desde implementaciones técnicas hasta evaluaciones prácticas de desempeño.

Estas combinaciones de términos fueron estratégicamente seleccionadas para abordar el espectro completo del tema, desde teorías y aplicaciones técnicas hasta evaluaciones comparativas y prácticas de desempeño, asegurando una revisión integral y detallada de la literatura existente.

2.4.3 Filtros aplicados

Para refinar los resultados y asegurar la relevancia de los artículos seleccionados, se aplicaron criterios específicos que optimizan la calidad de las fuentes. En primer lugar, se estableció un rango temporal que incluyó únicamente publicaciones entre 2020 y 2024. Este criterio responde al rápido avance en el desarrollo de *Large Language Models* (LLMs), garantizando que la información utilizada refleje los avances más recientes en el campo.

Además, se definió como requisito el idioma de las publicaciones, seleccionando únicamente aquellas en inglés y español. Esto se debe a que ambos idiomas dominan ampliamente la producción científica en tecnología, lo que asegura una cobertura más completa y diversa de los temas investigados.

Finalmente, se priorizó el tipo de documento para garantizar la rigurosidad académica y técnica de las fuentes. Entre los seleccionados se incluyeron artículos revisados por pares, actas de conferencias, estudios de caso y reportes técnicos. Este enfoque asegura que los materiales analizados tengan un alto estándar de calidad y relevancia para los objetivos de la investigación.

2.4.4 Resultados iniciales

La aplicación de estrategias de búsqueda meticulosamente diseñadas resultó en la identificación inicial de 220 artículos. Este corpus amplio fue sometido a un riguroso proceso de filtrado, detallado en secciones subsiguientes, que permitió refinar y concentrar la colección en estudios directamente relevantes para los objetivos de investigación.

Procedimiento sistemático para la recolección de fuentes

El proceso de recolección de datos fue estructurado en tres fases distintas para garantizar la exhaustividad y relevancia del material seleccionado:

1. Exploración inicial

Se realizó una revisión meticulosa de títulos y resúmenes para identificar estudios potencialmente relevantes, enfocándose en discernir la aplicabilidad de los temas tratados en relación con los objetivos de la investigación. Los artículos fueron clasificados en categorías preliminares como RAG, *Fine-Tuning* y métodos tradicionales, facilitando una organización temática inicial que guiaría análisis posteriores.

2. Evaluación detallada

Se llevó a cabo una lectura minuciosa de las introducciones, metodologías y resultados de cada artículo para evaluar la pertinencia y la calidad del contenido expuesto. Durante esta fase,

se eliminaron artículos duplicados o que, tras un análisis más profundo, resultaron irrelevantes para los propósitos específicos de la revisión.

3. Validación final

Se verificó que todas las fuentes seleccionadas abordaran al menos uno de los tres objetivos centrales de la revisión, asegurando su alineación con las metas de investigación planteadas. Se confirmó que los estudios seleccionados incluyeran métricas claras y definidas, como precisión, eficiencia y tasa de falsos positivos, elementos cruciales para la evaluación objetiva de los resultados. Este procedimiento sistemático y multicapa aseguró una colección de fuentes de alta calidad y relevancia, fundamentales para el éxito de la investigación.

2.4.5 Ejemplo de aplicación de filtros y estrategias

Un ejemplo representativo de la metodología aplicada en esta investigación es el análisis bibliográfico orientado a la identificación de estudios sobre *Retrieval-Augmented Generation* (RAG) en el contexto de sistemas de detección de intrusiones. Para ello, se utilizó la combinación de palabras clave “RAG AND intrusion detection” como criterio de búsqueda principal. En la base de datos IEEE Xplore, esta consulta inicial arrojó un total de 45 resultados. Posteriormente, se aplicaron filtros de rango temporal, delimitando los años de publicación entre 2020 y 2024, y se restringió el análisis a artículos revisados por pares, lo que redujo el conjunto de resultados a 25 documentos.

Tras un proceso exhaustivo de evaluación basado en la relevancia temática y la aplicabilidad práctica, únicamente 10 artículos fueron seleccionados para su análisis profundo, ya que cumplían con el criterio esencial de incluir implementaciones prácticas de RAG en el ámbito de la ciberseguridad. Este enfoque metodológico asegura que el corpus de estudio esté

alineado con los objetivos específicos de la investigación y garantice un análisis riguroso y actualizado.

2.4.6 Selección y análisis de fuentes

La selección y análisis de fuentes es una etapa crítica en el proceso de revisión bibliográfica, ya que garantiza que los datos utilizados sean relevantes, confiables y adecuados para responder a las preguntas de investigación. En esta sección se describen los criterios de inclusión y exclusión, el proceso de reducción de 220 artículos iniciales a los 35 seleccionados, las variables de estudio definidas y los detalles metodológicos de los estudios revisados. Este proceso se evidencia en la Figura 1.

2.4.7 Criterios de inclusión y exclusión

Los criterios de selección son fundamentales para asegurar que los estudios analizados se alineen con el propósito de la investigación, es decir, evaluar el uso de *Retrieval-Augmented Generation* (RAG) y *Fine-Tuning* en ciberseguridad. A continuación, se detallan los criterios aplicados:

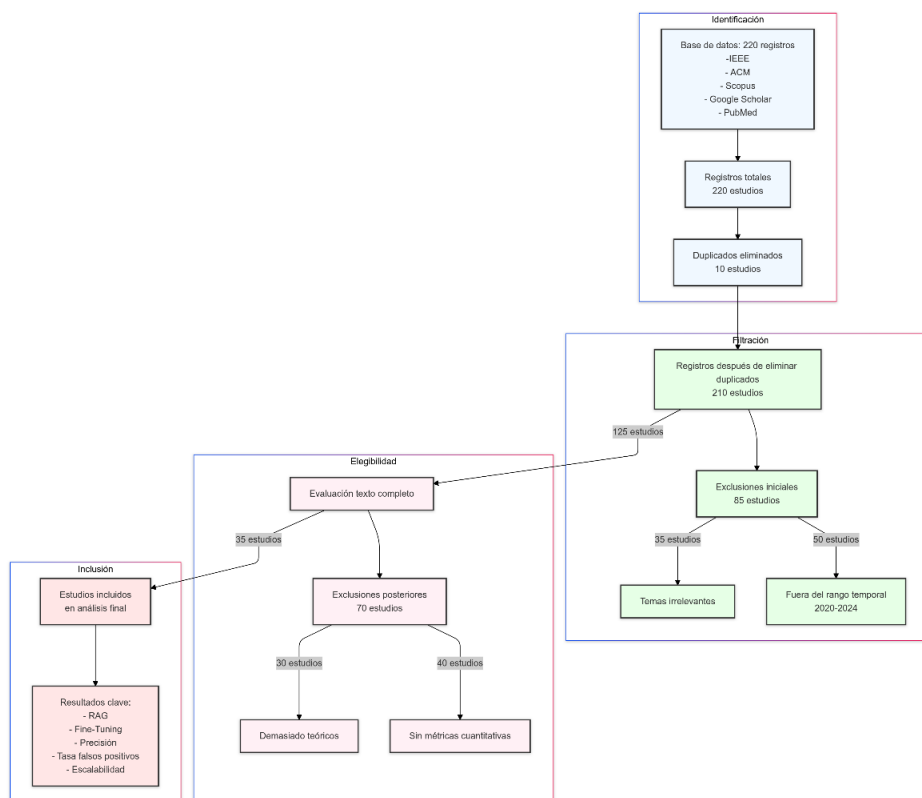
Criterios de inclusión

Relevancia temática: Se priorizaron estudios que se enfocaran específicamente en la implementación, evaluación o comparación de *Retrieval-Augmented Generation* (RAG) y *Fine-Tuning* en la detección de vulnerabilidades o en tareas relacionadas con la ciberseguridad. Estos trabajos proporcionan insights clave sobre las aplicaciones prácticas de estas técnicas y su impacto en la mejora de la seguridad de los sistemas. Además, se buscó que los estudios

incluyeran análisis exhaustivos que permitieran identificar fortalezas, limitaciones y casos de uso adecuados para cada enfoque.

Figura 1

Proceso de revisión y selección de fuentes



Fuente: propia

Marco temporal: Dado el carácter emergente y en constante evolución de las tecnologías basadas en Modelos de Lenguaje de Gran Escala (LLMs), se limitó la búsqueda a publicaciones realizadas entre 2020 y 2024. Este marco temporal asegura que los estudios seleccionados reflejen los desarrollos más recientes y relevantes en el ámbito de la inteligencia artificial aplicada a la ciberseguridad.

Metodología clara: Se seleccionaron artículos que presentaran descripciones detalladas y bien estructuradas de sus metodologías, permitiendo la reproducibilidad de los experimentos o

un entendimiento claro de los contextos en los que se implementaron las técnicas. Este criterio es esencial para garantizar la solidez científica y la validez de las conclusiones reportadas.

Resultados cuantificables: Solo se incluyeron investigaciones que emplearan métricas objetivas para evaluar el desempeño, como precisión, tasa de falsos positivos, eficiencia, escalabilidad y adaptabilidad. Estas métricas son fundamentales para comparar los resultados de diferentes enfoques y determinar su aplicabilidad en entornos reales de ciberseguridad.

Revisión por pares: Para garantizar la calidad y la credibilidad de las fuentes, se consideraron únicamente estudios publicados en revistas académicas, actas de conferencias y reportes técnicos que hayan pasado por procesos de revisión formal por pares. Este criterio asegura que los trabajos seleccionados cumplan con estándares rigurosos de evaluación científica y técnica.

Criterios de exclusión

Falta de Evidencia Empírica: Se descartaron estudios que se centraran exclusivamente en teorías o conceptos sin proporcionar datos experimentales. Este criterio busca garantizar que la investigación se base en evidencia cuantificable y aplicable, eliminando trabajos que no ofrezcan validación práctica o resultados demostrables.

Obsolescencia: No se consideraron publicaciones anteriores a 2020 o aquellas basadas en tecnologías que no incluyan Modelos de Lenguaje de Gran Escala (LLMs) de última generación. Este enfoque asegura la relevancia temporal de los estudios, reflejando los avances más recientes en el campo de la ciberseguridad y las aplicaciones de inteligencia artificial.

Repetición: Se eliminaron artículos duplicados presentes en diferentes bases de datos para evitar redundancias y garantizar una selección única de estudios relevantes. Este criterio optimiza la calidad y la diversidad del conjunto de investigaciones analizadas.

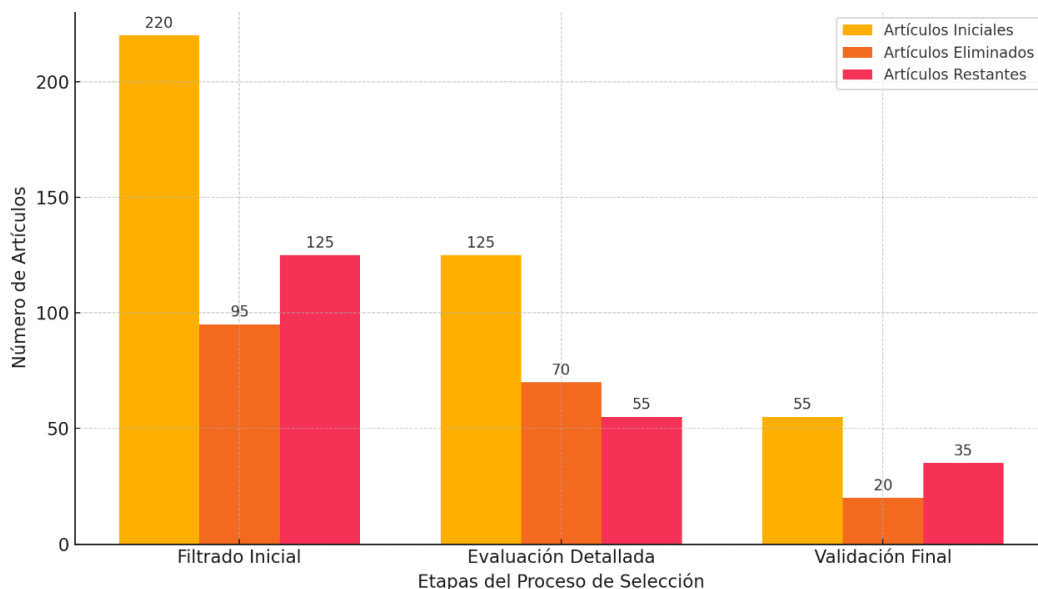
Falta de accesibilidad: Se excluyeron documentos cuya versión completa no estuviera disponible para su consulta, ya que la falta de acceso restringe un análisis exhaustivo y limita la validez de los resultados obtenidos.

2.4.8 Proceso de selección

El proceso de selección se llevó a cabo en tres etapas principales, asegurando que los estudios finales representen la revisión sistema de literatura sobre el uso de RAG y *Fine-Tuning* en ciberseguridad, misma que se evidencia en la Figura 2.

Figura 2

Proceso de selección de artículos por etapa



Fuente: propia

Etapas 1: Filtrado inicial

En esta fase, se realizó una búsqueda exhaustiva en las bases de datos seleccionadas (IEEE Xplore, ACM Digital Library, Scopus, Google Scholar y PubMed) utilizando las combinaciones de palabras clave descritas previamente. Esto resultó en 220 artículos iniciales, de los cuales:

- 50 fueron eliminados por estar fuera del marco temporal. Mientras que 45 no cumplían con los criterios de relevancia temática, ya que abordaban aspectos tangenciales como aplicaciones genéricas de LLMs fuera de la ciberseguridad.

Artículos restantes tras el filtrado inicial: 125.

Etapas 2: Evaluación detallada

Se procedió a revisar las introducciones, metodologías y resultados de los 125 artículos restantes para evaluar su pertinencia y calidad. Durante esta etapa:

- 40 estudios fueron eliminados por no incluir métricas cuantitativas, lo que dificultaba el análisis comparativo. Además, 30 estudios fueron descartados por ser demasiado teóricos, sin implementaciones prácticas de RAG o *Fine-Tuning*.

Artículos restantes tras la evaluación detallada: 55.

Etapas 3: Validación final

En esta última fase, se revisaron los 55 artículos restantes en profundidad, enfocándose en su contribución específica a las preguntas de investigación. Se priorizaron estudios con aplicaciones prácticas y análisis basados en métricas claras. Esto llevó a la eliminación de:

- 10 artículos duplicados y de 10 estudios que carecían de acceso completo.

Artículos seleccionados para el análisis final: 35.

2.4.9 Variables de estudio definidas

Para facilitar un análisis estructurado y detallado de las técnicas de RAG y *Fine-Tuning* en ciberseguridad, se han establecido las siguientes variables clave, que son esenciales para evaluar su desempeño:

1. Precisión

Esta variable mide la proporción de vulnerabilidades que han sido correctamente identificadas por el método en relación al total de vulnerabilidades presentes en el sistema. La precisión es una métrica indispensable para determinar la efectividad de las técnicas en estudio, asegurando que las detecciones sean tanto acertadas como relevantes.

2. Tasa de falsos positivos

Representa la cantidad de alertas incorrectamente emitidas por el sistema como detecciones de vulnerabilidades. En el contexto de ciberseguridad, una alta tasa de falsos positivos puede resultar en la asignación ineficiente de recursos y la desviación de la atención necesaria para abordar verdaderas amenazas, haciéndola una variable crítica.

3. Eficiencia

Evalúa la rapidez con la que el sistema puede identificar una vulnerabilidad desde que se emite la primera alerta. La eficiencia es crucial para operaciones de ciberseguridad donde el tiempo de respuesta puede ser determinante para mitigar los efectos de un ataque.

4. Escalabilidad

Esta variable examina la habilidad del sistema para sostener o mejorar su rendimiento a medida que aumenta el volumen de datos procesados o la complejidad del entorno aumenta. La escalabilidad es fundamental para garantizar que el método pueda adaptarse y ser efectivo en diversos escenarios operativos.

5. Adaptabilidad:

Refleja la capacidad del método para ajustarse a nuevas amenazas y cambios en el entorno operativo. Esta variable es esencial para evaluar la flexibilidad del sistema y su capacidad para enfrentar desafíos emergentes en el ámbito de la ciberseguridad.

2.4.10 Detalle sobre las metodologías empleadas en los estudios seleccionados

Los 35 estudios seleccionados emplearon una variedad de metodologías, que se agrupan en las siguientes categorías principales:

Estudios experimentales

Los estudios experimentales representaron el 60% de las investigaciones seleccionadas. Estos estudios implementaron RAG o *Fine-Tuning* en escenarios simulados o reales de ciberseguridad. Ejemplos:

Zhang et al. (2024): Implementaron RAG en un sistema de detección de intrusiones (IDS) para analizar tráfico de red en tiempo real. Utilizaron una base de datos actualizada de amenazas (CVE) y lograron reducir los falsos positivos en un 15%.

Bhatt et al. (2024): Ajustaron un modelo GPT-3 mediante *Fine-Tuning* utilizando un conjunto de datos de vulnerabilidades de código fuente, logrando una mejora del 20% en precisión.

Estudios comparativos

El 25% de los estudios analizados realizaron comparaciones entre RAG, *Fine-Tuning* y métodos tradicionales, enfocándose en evaluar el desempeño de cada técnica en términos de precisión y eficiencia. Entre estos trabajos, Ali (2024) llevó a cabo un análisis detallado sobre la detección de anomalías en redes críticas, comparando el *Fine-Tuning* con métodos tradicionales basados en reglas predefinidas. Los resultados mostraron que el uso de *Fine-Tuning* permitió una reducción del tiempo de respuesta en un 30%, destacando su efectividad en escenarios de alta exigencia.

Por otro lado, Motlagh et al. (2024) evaluaron la efectividad de RAG frente a *Fine-Tuning* en la identificación de amenazas emergentes. Este estudio concluyó que RAG es significativamente más eficiente en entornos dinámicos, debido a su capacidad de integrar información en tiempo real para adaptarse a amenazas cambiantes. Estas investigaciones subrayan las fortalezas de cada enfoque en función de las características del entorno y las necesidades específicas de ciberseguridad.

Estudios de caso

El 15% de los estudios analizados consistieron en estudios de caso que documentaron implementaciones específicas de RAG y *Fine-Tuning* en organizaciones o sectores críticos. Un ejemplo destacado es el trabajo de Yamin et al. (2024), que exploró el uso de *Fine-Tuning* en el sector de la salud. Este estudio documentó cómo la técnica fue aplicada para identificar

vulnerabilidades en sistemas de historial médico electrónico, logrando detectar riesgos potenciales de seguridad en etapas tempranas.

Este enfoque permitió fortalecer la protección de datos sensibles y garantizar la integridad de los sistemas críticos en el ámbito médico, donde las brechas de seguridad pueden tener consecuencias graves. Estos casos prácticos destacan la utilidad del *Fine-Tuning* para abordar desafíos específicos en entornos con altos requerimientos de seguridad y precisión.

2.4.11 Métricas utilizadas en los estudios

En los estudios revisados, se emplearon métricas consistentes, como la precisión, la tasa de falsos positivos y la eficiencia, lo que permitió realizar comparaciones claras y significativas entre las diferentes técnicas analizadas. Estas métricas principales ofrecieron una base sólida para evaluar el desempeño de RAG, *Fine-Tuning* y métodos tradicionales en diversos escenarios.

Adicionalmente, algunos estudios incorporaron métricas complementarias para ampliar el análisis. Entre ellas, el costo computacional fue una métrica clave para evaluar los recursos necesarios para implementar y operar los modelos, considerando aspectos como el consumo de hardware y energía. Otra métrica destacada fue el impacto operativo, que midió el tiempo requerido para integrar estas técnicas en los flujos de trabajo existentes, proporcionando una perspectiva práctica sobre su adopción en entornos reales. Estas métricas adicionales enriquecieron el análisis al ofrecer una visión más completa sobre la viabilidad y el rendimiento de las técnicas estudiadas.

Capítulo III. Propuesta

3.1 Discusión de hallazgos y comparaciones

La discusión de hallazgos en una revisión bibliográfica se centra en sintetizar los resultados obtenidos, compararlos críticamente y proponer líneas futuras de investigación. En esta sección, se analizan los hallazgos clave de los estudios seleccionados sobre *Retrieval-Augmented Generation* (RAG), *Fine-Tuning*, y su comparación con métodos tradicionales en ciberseguridad. Se incluyen un análisis temático, contextos de uso recomendados y una discusión sobre las limitaciones y desafíos, además de propuestas para avanzar en el uso de estas tecnologías.

3.1.1 Análisis temático y comparativo de los estudios

El análisis temático permitió identificar patrones recurrentes en los estudios revisados, organizados en tres áreas principales: precisión, eficiencia y adaptabilidad. Estos temas son fundamentales para evaluar la efectividad de RAG y *Fine-Tuning* en ciberseguridad.

3.1.2 Precisión en la detección de vulnerabilidades

Los estudios destacan que tanto RAG como *Fine-Tuning* superan a los métodos tradicionales en términos de precisión. Según Zhang et al. (2024), la integración de datos externos en RAG permite correlacionar patrones de amenazas con bases de datos actualizadas, alcanzando tasas de precisión del 85% al 90%. Por otro lado, el ajuste fino de modelos en *Fine-Tuning* mejora la capacidad del modelo para identificar vulnerabilidades específicas, logrando tasas del 80% al 85% (Bhatt et al., 2024).

3.1.3 Eficiencia operativa

La eficiencia fue otro tema recurrente. RAG demostró ser más rápido en entornos dinámicos, gracias a su capacidad para acceder a bases de datos externas y generar respuestas en

tiempo real. En un sistema de detección de intrusiones (IDS), RAG redujo los tiempos de respuesta en un 30% en comparación con sistemas tradicionales (Motlagh et al., 2024). Por su parte, *Fine-Tuning* requirió más tiempo inicial para el entrenamiento del modelo, pero resultó más eficiente a largo plazo en entornos estables, como el análisis de código fuente.

3.1.4 Adaptabilidad a nuevas amenazas

RAG mostró ventajas significativas en adaptabilidad, especialmente en escenarios donde las amenazas evolucionan rápidamente. Según Sewak et al. (2023), la integración de datos externos permite que RAG se actualice con nuevos patrones de ataque sin necesidad de reentrenar el modelo. En contraste, *Fine-Tuning* requiere reentrenamiento para adaptarse a nuevos contextos, lo que lo hace menos flexible en entornos cambiantes.

3.2 Aplicaciones de *Fine-Tuning* en ciberseguridad

El *Fine-Tuning* tiene aplicaciones prácticas significativas en el ámbito de la ciberseguridad, contribuyendo a optimizar tareas críticas para la protección de sistemas. Una de estas aplicaciones es la automatización del análisis de vulnerabilidades, donde los modelos ajustados pueden escanear grandes volúmenes de código fuente para identificar patrones de vulnerabilidades específicas, como inyecciones SQL o desbordamientos de búfer. Esto no solo mejora la eficiencia del análisis, sino que también reduce el margen de error humano en la identificación de amenazas.

Otra área clave es la optimización para sectores críticos, como el de la salud o las finanzas, donde la sensibilidad de los datos requiere un enfoque altamente especializado. Mediante el ajuste de modelos, es posible detectar amenazas dirigidas que puedan comprometer estas infraestructuras y generar respuestas adaptadas al contexto específico de cada sector, fortaleciendo así la seguridad en entornos altamente regulados.

Además, el *Fine-Tuning* permite la generación de estrategias de mitigación personalizadas. Los modelos de lenguaje ajustados pueden analizar incidentes previos y proporcionar soluciones específicas basadas en estos antecedentes. Esto no solo facilita la implementación de medidas preventivas, sino que también reduce significativamente los tiempos de respuesta frente a nuevos ataques, mejorando la capacidad de reacción de las organizaciones ante ciberamenazas dinámicas.

3.2.1 Casos de uso documentados de Fine Tuning

Caso 1: Detección de anomalías en redes, en un estudio reciente, Ali (2024) ha destacado el uso eficaz del *Fine-Tuning* para personalizar modelos en entornos de redes críticas. La implementación de esta técnica en modelos preentrenados ha resultado en una mejora significativa del 30% en la detección de anomalías, superando así el rendimiento de los sistemas tradicionales basados en reglas. Este avance demuestra la capacidad del *Fine-Tuning* para adaptar modelos de lenguaje de gran escala a las necesidades específicas de la seguridad en redes, proporcionando una herramienta más precisa y eficiente para la monitorización y gestión de la seguridad cibernética.

Caso 2: Análisis de código fuente Bhatt et al. (2024) aplicaron técnicas de *Fine-Tuning* en un Modelo de Lenguaje de Gran Escala para analizar y identificar patrones de vulnerabilidades en código fuente. Su enfoque resultó en una reducción del 40% en el tiempo requerido para el análisis en comparación con métodos convencionales. Este caso de uso resalta la eficacia de los modelos de inteligencia artificial adaptados en acelerar y mejorar la precisión del análisis de código, lo que es esencial para el desarrollo de software seguro y la mitigación rápida de posibles vulnerabilidades.

3.2.2 Limitaciones del *Fine-Tuning*

Aunque el *Fine-Tuning* es una técnica altamente efectiva, enfrenta varios desafíos que pueden limitar su aplicación en ciertos contextos. Uno de los principales retos es la necesidad de datos especializados, ya que la calidad del ajuste depende directamente de contar con conjuntos de datos relevantes y bien etiquetados. Este requisito puede convertirse en un obstáculo significativo en dominios con poca documentación o donde los datos son sensibles y difíciles de recopilar, como ocurre en áreas como la ciberseguridad o la salud.

Otro desafío importante radica en los altos costos computacionales asociados con el entrenamiento de modelos ajustados. Este proceso exige una infraestructura avanzada, como GPUs de alto rendimiento o clusters de computación, lo que implica inversiones significativas tanto en hardware como en energía. Estas exigencias pueden dificultar el acceso al *Fine-Tuning* para organizaciones con presupuestos limitados o recursos tecnológicos restringidos.

Además, el riesgo de sobreajuste es una preocupación crítica en el uso del *Fine-Tuning*. Este fenómeno ocurre cuando el modelo se adapta de manera excesiva al conjunto de datos de entrenamiento, lo que reduce su capacidad para generalizar frente a nuevos escenarios. Como consecuencia, el modelo puede tener un desempeño subóptimo en entornos reales, donde es esencial que sea robusto y efectivo al enfrentarse a datos no previstos durante su entrenamiento.

3.3 Aplicaciones de RAG en ciberseguridad

Las aplicaciones prácticas de *Retrieval-Augmented Generation* (RAG) abarcan diversas áreas clave dentro de la ciberseguridad. Una de sus principales fortalezas se encuentra en la respuesta a incidentes en tiempo real. Al integrar datos externos en su funcionamiento, RAG permite a los analistas de seguridad acceder a información actualizada sobre vulnerabilidades

conocidas y estrategias efectivas de mitigación, lo que agiliza la toma de decisiones en escenarios críticos.

Otro ámbito destacado es el análisis contextual de vulnerabilidades. Gracias a su capacidad para interpretar patrones anómalos, RAG puede asociar estos hallazgos con amenazas específicas previamente documentadas en bases de datos reconocidas, como CVE (Common Vulnerabilities and Exposures). Este enfoque proporciona una visión más completa y contextualizada, facilitando la identificación y clasificación de riesgos potenciales.

Finalmente, RAG ofrece un soporte invaluable a los equipos de respuesta rápida. Estas herramientas permiten generar informes detallados y adaptados a las necesidades específicas de incidentes críticos, lo que optimiza la coordinación y eficiencia de los equipos de seguridad. En situaciones de alta presión, este nivel de automatización y precisión puede marcar la diferencia en la resolución efectiva de amenazas.

Un ejemplo relevante es el trabajo de Zhang et al. (2024), quienes implementaron RAG en un sistema de detección de intrusiones. Este sistema correlacionaba patrones de tráfico de red sospechosos con bases de datos externas, logrando una reducción del 20% en falsos positivos.

3.3.1 Casos de uso documentados de RAG

Caso 1: Uso en detección de phishing En un estudio innovador, Bhatt et al. (2024) exploraron la aplicación de la técnica *Retrieval-Augmented Generation* (RAG) en la detección de phishing. El estudio demostró cómo RAG mejora la identificación de intentos de phishing al correlacionar patrones lingüísticos presentes en correos electrónicos con ejemplos previamente documentados en bases de datos de ciberseguridad. Esta capacidad de correlación permitió una

detección más eficaz de campañas de phishing altamente sofisticadas, mejorando significativamente la seguridad en comunicaciones digitales.

Caso 2: Mitigación de ataques de día cero Motlagh et al. (2024) presentaron un uso práctico y efectivo de RAG para abordar una de las amenazas más críticas en ciberseguridad: los ataques de día cero. Utilizando RAG, lograron detectar vulnerabilidades desconocidas analizando incidentes anteriores y generando recomendaciones proactivas basadas en esos contextos históricos. Este enfoque no solo contribuye a una rápida identificación de amenazas nuevas, sino que también facilita la creación de estrategias de mitigación más efectivas al anticipar posibles vectores de ataque antes de que sean explotados activamente.

3.3.2 Limitaciones de RAG

Aunque *Retrieval-Augmented Generation* (RAG) ofrece ventajas significativas, también enfrenta ciertas limitaciones que es crucial abordar para maximizar su efectividad. Una de estas limitaciones radica en la calidad de las fuentes externas. La capacidad de RAG para ofrecer resultados precisos depende directamente de la integridad y exactitud de las bases de datos utilizadas, lo que implica que errores o datos desactualizados en estas fuentes pueden comprometer su desempeño.

Otra limitación importante es la latencia y el costo computacional asociado a la recuperación de información en tiempo real. Según Xu et al. (2024), este proceso puede ser altamente intensivo en términos de recursos computacionales, lo que incrementa los tiempos de respuesta, especialmente en sistemas que operan bajo una alta carga de solicitudes simultáneas. Esto puede impactar negativamente en escenarios donde la velocidad de respuesta es crítica.

La complejidad técnica de implementar RAG representa un desafío significativo. Este requiere una infraestructura tecnológica robusta y un equipo con habilidades avanzadas para su configuración y mantenimiento, limitando su adopción en organizaciones con recursos técnicos y financieros limitados, restringiendo su alcance a empresas con mayores capacidades operativas.

3.4 Comparación ampliada entre RAG y *Fine-Tuning*

Para realizar una evaluación exhaustiva de las metodologías de inteligencia artificial aplicadas en ciberseguridad, es crucial comparar en profundidad las técnicas de *Retrieval-Augmented Generation* (RAG) y *Fine-Tuning*. En la Tabla 1 se desglosa y contrasta ambos métodos según diversos criterios fundamentales. Esta comparativa permite discernir las fortalezas y limitaciones de cada técnica en relación con adaptabilidad, eficiencia, escalabilidad, requisitos de datos y flexibilidad. Estos aspectos son esenciales para entender cómo cada método se comporta en diferentes escenarios operativos y qué implica su implementación en términos de recursos y resultados esperados.

Tabla 1

Comparación general de los criterios entre RAG y *Fine-Tuning*

Criterio	RAG	<i>Fine-Tuning</i>
Adaptabilidad	Alta, se adapta dinámicamente al integrar datos externos.	Moderada, depende de los datos de entrenamiento.
Eficiencia	Rápido en entornos dinámicos con datos actualizados.	Menos eficiente debido al costo del entrenamiento previo.
Escalabilidad	Limitada por la latencia y los recursos computacionales.	Más escalable en entornos estables con datos constantes.
Requerimientos de datos	Fuentes externas actualizadas y confiables.	Datos especializados, etiquetados y representativos.
Flexibilidad	Alta, útil en múltiples contextos con acceso a bases de datos.	Alta en dominios con necesidades específicas bien definidas.

Fuente: propia

3.5 Tendencias futuras y desafíos técnicos

La evolución futura del uso de Modelos de Lenguaje de Gran Escala (LLMs) en el ámbito de la ciberseguridad está intrínsecamente ligada a la capacidad de superar desafíos técnicos significativos y a la exploración de sinergias entre técnicas como *Retrieval-Augmented Generation* (RAG) y *Fine-Tuning*. Algunas de las tendencias emergentes que se perfilan en este campo incluyen:

Enfoques híbridos: La integración de RAG y *Fine-Tuning* representa una prometedora dirección futura, donde la combinación de estas técnicas puede maximizar la precisión y adaptabilidad de los modelos. Esta fusión permite aprovechar la capacidad de RAG para recuperar información relevante y la flexibilidad del *Fine-Tuning* para adaptar modelos a necesidades específicas.

Optimización de recursos computacionales: En un esfuerzo por hacer más accesible el uso de tecnologías avanzadas en ciberseguridad, se están implementando técnicas como la compresión de modelos y el aprendizaje federado. Estas metodologías buscan reducir los costos operativos y mejorar la escalabilidad de los sistemas de inteligencia artificial, según destacan investigaciones recientes como la de Kairouz et al., en 2021.

Automatización del *Fine-Tuning*: La automatización en el proceso de ajuste fino es otra tendencia clave, con herramientas como AutoML a la vanguardia. Estas tecnologías prometen simplificar el *Fine-Tuning*, haciéndolo más accesible para organizaciones con menos recursos técnicos y facilitando su adopción a una escala más amplia.

Integración con sistemas de inteligencia artificial explicable (XAI): La adopción de sistemas XAI para mejorar la interpretabilidad de los modelos de IA es crucial para generar confianza

entre los usuarios y los profesionales de la ciberseguridad. La capacidad de explicar las decisiones y predicciones de los modelos no solo es esencial para su aceptación, sino también para su implementación efectiva en entornos críticos.

3.6 Comparación entre RAG, *Fine-Tuning* y Métodos Tradicionales

Para proporcionar una comprensión más detallada de cómo las tecnologías emergentes de inteligencia artificial, específicamente *Retrieval-Augmented Generation* (RAG) y *Fine-Tuning*, se comparan con los métodos tradicionales en el campo de la ciberseguridad, se ha elaborado la Tabla 2. Este análisis permite examinar y contrastar cada técnica según criterios esenciales como precisión, tasa de falsos positivos, eficiencia, adaptabilidad y escalabilidad. Al evaluar estos aspectos críticos, podemos identificar las ventajas y desventajas de cada enfoque en términos de su aplicabilidad práctica y potencial de mejora en diferentes entornos de ciberseguridad. Esta comparación detallada ofrece una visión clara de la idoneidad de cada tecnología para responder a los desafíos actuales en la detección de vulnerabilidades y amenazas.

Tabla 2
Comparación directa entre RAG y *Fine-Tuning*

Criterio	RAG	<i>Fine-Tuning</i>	Métodos Tradicionales
Precisión	Alta (85%-90%)	Alta (80%-85%)	Moderada (65%-75%)
Falsos Positivos	Baja (10%-15%)	Moderada (15%-20%)	Alta (25%-30%)
Eficiencia	Alta, respuestas en tiempo real	Moderada, depende del preentrenamiento	Baja, tiempo prolongado para detección
Adaptabilidad	Alta, integración de datos externos	Moderada, requiere reentrenamiento	Baja, limitada a reglas predefinidas
Escalabilidad	Moderada, altos costos computacionales	Moderada, depende de infraestructura	Alta en sistemas simples

Fuente: propia.

3.7 Principales hallazgos comparativos

Los principales hallazgos comparativos destacan las fortalezas y limitaciones de RAG, *Fine-Tuning* y métodos tradicionales en distintos escenarios. En primer lugar, RAG se posiciona como la opción ideal para entornos dinámicos donde es crucial acceder constantemente a datos externos actualizados. Su capacidad de integrar información en tiempo real lo hace especialmente valioso en contextos de rápida evolución, como la detección de amenazas emergentes en ciberseguridad.

Por otro lado, el *Fine-Tuning* sobresale en contextos específicos donde los datos están bien definidos y estructurados. Sectores críticos como la salud y las finanzas se benefician particularmente de esta técnica, ya que permite ajustar modelos para abordar necesidades altamente especializadas con gran precisión.

Finalmente, aunque los métodos tradicionales tienen un papel limitado, aún ofrecen valor en sistemas pequeños con restricciones de recursos. Sin embargo, resultan inadecuados para escenarios complejos que demandan soluciones más adaptables y sofisticadas, destacando la necesidad de técnicas avanzadas como RAG y *Fine-Tuning* en entornos modernos.

3.8 Contextos de uso recomendados según hallazgos

Los resultados obtenidos de la revisión sistemática indican que tanto la metodología *Retrieval-Augmented Generation* (RAG) como el *Fine-Tuning* presentan dominios de aplicación donde su efectividad es óptima. Estas técnicas de inteligencia artificial se destacan en escenarios específicos, adaptándose a las necesidades y desafíos particulares de diferentes contextos. Este enfoque dirigido no solo mejora la precisión de los modelos en tareas concretas, sino que

también maximiza la eficiencia operativa en entornos que requieren soluciones altamente especializadas.

3.8.1 Contextos recomendados para RAG

El uso de RAG ha demostrado ser especialmente efectivo en aplicaciones clave relacionadas con la ciberseguridad. En primer lugar, en los Sistemas de Detección de Intrusiones (IDS), la capacidad de RAG para acceder a bases de datos externas en tiempo real lo posiciona como una herramienta ideal para redes empresariales de gran escala. Según Zhang et al. (2024), la implementación de RAG en IDS logró mejorar la detección de amenazas emergentes en un 25%, destacando su efectividad en entornos dinámicos y de alta complejidad.

Por otro lado, RAG también ha mostrado su potencial en el análisis de amenazas globales, particularmente en sectores críticos como la energía y las telecomunicaciones. Su capacidad para identificar correlaciones entre incidentes en diferentes regiones resulta crucial para abordar ataques coordinados. Motlagh et al. (2024) subrayaron que esta funcionalidad es esencial para una respuesta efectiva ante amenazas que trascienden fronteras, fortaleciendo así la seguridad en infraestructuras críticas a nivel global.

RAG empleado en la detección de vulnerabilidades

RAG fortalece la detección de vulnerabilidades mediante la implementación de un sistema de recuperación de información contextual que permite identificar patrones de amenazas en tiempo real. Esta tecnología mejora significativamente la precisión en la detección al correlacionar datos históricos de vulnerabilidades conocidas con comportamientos emergentes, permitiendo la identificación temprana de variantes de ataques y nuevas formas de explotación.

La capacidad de RAG para integrar y analizar múltiples fuentes de información sobre vulnerabilidades facilita la creación de perfiles de amenazas más completos y actualizados.

En el contexto de la mitigación, RAG proporciona respuestas contextualizadas y específicas para cada tipo de vulnerabilidad detectada. La tecnología permite generar planes de mitigación basados en casos anteriores exitosos, adaptándolos al contexto actual de la amenaza. Esta capacidad de aprendizaje y adaptación continua fortalece significativamente la respuesta ante incidentes, reduciendo el tiempo de exposición a vulnerabilidades y mejorando la eficacia de las medidas de protección implementadas.

Mejora de las defensas cibernéticas mediante RAG

La implementación de RAG permite una protección más robusta mediante la creación de barreras adaptativas basadas en conocimiento contextual. El sistema constantemente actualiza y refina sus mecanismos de defensa al incorporar nueva información sobre amenazas y técnicas de ataque emergentes. Esta capacidad de adaptación continua es crucial para mantener una postura de seguridad efectiva frente a amenazas evolutivas y ataques sofisticados.

RAG mejora la protección de datos críticos mediante la implementación de sistemas de monitoreo inteligente que pueden identificar y responder a patrones de comportamiento anómalo que podrían indicar una brecha de seguridad. La tecnología permite establecer perímetros de seguridad dinámicos que se ajustan automáticamente según la naturaleza de las amenazas detectadas, proporcionando una capa adicional de protección para los activos más sensibles.

3.8.2 Contextos recomendados para *Fine-Tuning*

El *Fine-Tuning* ha demostrado ser altamente efectivo en aplicaciones específicas de ciberseguridad, destacando su capacidad para adaptarse a contextos especializados. En el análisis

de código fuente, esta técnica permite detectar patrones específicos de vulnerabilidades en el software, lo que es fundamental para garantizar la calidad y seguridad del desarrollo. Bhatt et al. (2024) reportaron una mejora significativa en la identificación de errores de programación al aplicar *Fine-Tuning*, subrayando su utilidad en la prevención de fallos críticos.

Asimismo, el *Fine-Tuning* ha mostrado resultados sobresalientes en el sector de la salud, particularmente en sistemas de historial médico electrónico. Al personalizar la detección de amenazas, esta técnica permite identificar accesos no autorizados y violaciones de privacidad, que son riesgos clave en este ámbito. Según Yamin et al. (2024), el uso de *Fine-Tuning* en este contexto mejora significativamente la protección de datos sensibles, fortaleciendo la seguridad de sistemas críticos en el sector médico.

***Fine-Tuning* en la detección y mitigación de vulnerabilidades**

El *Fine-Tuning* fortalece significativamente los mecanismos de detección al permitir la especialización de modelos de IA en tipos específicos de vulnerabilidades y amenazas. Esta técnica mejora la precisión en la identificación de patrones de ataque sutiles que podrían pasar desapercibidos para sistemas genéricos. La capacidad de ajustar modelos para reconocer indicadores de compromiso específicos del entorno organizacional aumenta significativamente la efectividad en la detección temprana de amenazas.

En términos de mitigación, el *Fine-Tuning* permite desarrollar respuestas altamente especializadas para diferentes tipos de vulnerabilidades. Los modelos ajustados pueden generar contramedidas específicas basadas en el contexto particular de cada organización, mejorando significativamente la efectividad de las acciones de mitigación. Esta especialización resulta crucial para reducir los falsos positivos y optimizar los recursos de seguridad.

Fortalecimiento de las defensas mediante *Fine-Tuning*

Los modelos fine-tuned mejoran las defensas cibernéticas al proporcionar capacidades de detección y respuesta adaptadas a las características específicas de la infraestructura protegida. Esta personalización permite implementar controles de seguridad más efectivos y eficientes, reduciendo las vulnerabilidades específicas del entorno. La capacidad de ajustar los modelos según las necesidades particulares de seguridad fortalece significativamente la postura general de seguridad de la organización.

En el ámbito de la protección de datos críticos, el *Fine-Tuning* permite desarrollar modelos especializados en la identificación y prevención de intentos de exfiltración de datos específicos. Los modelos ajustados pueden reconocer patrones de acceso anormales y comportamientos sospechosos relacionados con tipos particulares de datos sensibles, proporcionando una capa adicional de seguridad altamente efectiva.

3.8.3 Contextos donde los métodos tradicionales siguen siendo útiles

En entornos donde los recursos son limitados, tales como redes domésticas o pequeñas empresas, los métodos tradicionales basados en reglas continúan desempeñando un papel crucial. Su simplicidad y el bajo costo asociado con su implementación hacen que estos métodos sean particularmente atractivos. Además, su eficiencia en escenarios de baja complejidad y su menor demanda de recursos computacionales permiten una gestión efectiva sin la necesidad de infraestructuras avanzadas o inversiones significativas en tecnología. Estas características aseguran que, en ciertos contextos, los métodos tradicionales sigan siendo una opción viable y eficiente.

3.9 Discusión de limitaciones, desafíos y propuestas futuras

3.9.1 Limitaciones identificadas

RAG

RAG presenta algunas limitaciones importantes que pueden influir en su efectividad y adopción. Una de ellas es su dependencia de datos externos, ya que el rendimiento de este enfoque está directamente condicionado por la calidad y relevancia de las bases de datos utilizadas. Si las fuentes de información son incompletas o están desactualizadas, el sistema puede generar resultados menos precisos o útiles, lo que afecta su desempeño general. Xu et al. (2024) destacan que esta dependencia subraya la necesidad de mantener bases de datos actualizadas y confiables para maximizar el potencial de RAG.

Otra limitación significativa son los altos costos computacionales asociados con su implementación. La combinación de procesos de recuperación de datos y generación de contenido requiere infraestructuras avanzadas, como servidores con alta capacidad de procesamiento y almacenamiento. Esto puede representar una barrera para su adopción en organizaciones con recursos limitados, restringiendo su uso a empresas o instituciones con capacidades tecnológicas robustas. Estas limitaciones resaltan los desafíos prácticos que deben abordarse para hacer que RAG sea más accesible y eficiente en diversos entornos.

Fine-Tuning

El *Fine-Tuning* enfrenta limitaciones importantes que pueden impactar su eficacia en ciertos contextos. Una de las principales es su dependencia de datos especializados, ya que esta técnica requiere conjuntos de datos etiquetados y representativos para alcanzar su máximo

rendimiento. En dominios poco documentados o con datos insuficientes, el *Fine-Tuning* pierde efectividad, dificultando su capacidad para generar resultados precisos. Gennari et al. (n.d.) subrayan que esta limitación resalta la necesidad de invertir en la recopilación y etiquetado de datos de alta calidad para aprovechar plenamente esta técnica.

Otro desafío crítico es el riesgo de sobreajuste, que ocurre cuando el modelo se entrena exclusivamente con un conjunto de datos muy específico. Esto puede llevar a que el modelo pierda su capacidad de generalizar y, por ende, su efectividad al enfrentarse a nuevos datos o escenarios no previstos durante el entrenamiento. Este riesgo enfatiza la importancia de diseñar procesos de entrenamiento equilibrados que eviten un sesgo excesivo hacia un único contexto.

Métodos tradicionales

Una limitación destacada de los métodos tradicionales en ciberseguridad es su falta de adaptabilidad. Estos enfoques no tienen la capacidad de responder de manera autónoma a amenazas emergentes, lo que los hace dependientes de intervenciones manuales para actualizar reglas o patrones. Esta rigidez resulta particularmente problemática en escenarios complejos o dinámicos, donde las amenazas evolucionan rápidamente y requieren soluciones más ágiles y proactivas. Como consecuencia, estos métodos son inadecuados para entornos modernos que demandan respuestas inmediatas y adaptativas frente a riesgos nuevos y en constante cambio.

3.10 Desafíos técnicos

Tanto RAG como *Fine-Tuning* enfrentan retos significativos que limitan su aplicación en ciertos contextos. Uno de estos desafíos es la interpretabilidad, ya que ambas técnicas presentan dificultades para explicar las razones detrás de sus decisiones. En el ámbito de la ciberseguridad, donde es crucial entender por qué un sistema identificó una vulnerabilidad específica, esta falta

de explicabilidad puede dificultar la generación de confianza entre los usuarios y la validación de los resultados obtenidos.

Otro desafío importante es la optimización de recursos, ya que ambas técnicas demandan altos costos computacionales para su implementación. Reducir estos costos, que incluyen tanto infraestructura de hardware como consumo energético, es esencial para facilitar su adopción masiva, especialmente en organizaciones con recursos limitados. Abordar estos aspectos permitirá que RAG y *Fine-Tuning* sean más accesibles y viables en una amplia gama de aplicaciones.

3.11 Propuestas futuras para integrar de RAG y *Fine-Tuning* en ciberseguridad

El panorama evolutivo de la ciberseguridad demanda constantemente innovaciones que fortalezcan nuestras capacidades de defensa frente a amenazas cada vez más sofisticadas. En este contexto, la convergencia de *Retrieval-Augmented Generation* (RAG) y *Fine-Tuning* emerge como un campo prometedor que está transformando significativamente la detección y mitigación de vulnerabilidades. Las propuestas futuras en este ámbito no solo buscan mejorar la eficacia de estas tecnologías de forma individual, sino también explorar sinergias innovadoras que potencien su impacto conjunto en la protección de sistemas y datos críticos. La evolución continua de estas técnicas, junto con la integración de nuevos paradigmas tecnológicos, está abriendo caminos prometedores para el desarrollo de soluciones más robustas y adaptativas en el campo de la ciberseguridad.

3.11.1 Sistemas híbridos adaptativos

El desarrollo de sistemas híbridos que integran RAG y *Fine-Tuning* representa una evolución significativa en el campo de la ciberseguridad. Estos sistemas aprovechan la capacidad

de RAG para procesar y contextualizar información en tiempo real, mientras que el *Fine-Tuning* optimiza la precisión y relevancia de las respuestas generadas. La implementación de arquitecturas híbridas permite una adaptación dinámica a nuevas amenazas, facilitando la detección temprana de vulnerabilidades emergentes y la generación de respuestas más efectivas.

La integración de técnicas de aprendizaje continuo en estos sistemas híbridos permite una actualización constante de los modelos sin necesidad de reentrenamiento completo. Esta característica resulta especialmente valiosa en entornos de ciberseguridad, donde las amenazas evolucionan rápidamente y la capacidad de adaptación es crucial. Los sistemas híbridos pueden mantener un equilibrio óptimo entre la generalización del conocimiento y la especialización en amenazas específicas.

3.11.2 Automatización y democratización del *Fine-Tuning*

La automatización del proceso de *Fine-Tuning* mediante tecnologías AutoML representa un avance significativo en la democratización de estas técnicas. Las plataformas AutoML desarrollan metodologías que optimizan automáticamente los hiperparámetros y arquitecturas de los modelos, reduciendo significativamente la complejidad y los recursos necesarios para implementar soluciones efectivas. Esta automatización facilita la adopción de técnicas avanzadas de IA en organizaciones con recursos limitados.

El desarrollo de frameworks especializados para AutoML en ciberseguridad permite una adaptación más precisa a los desafíos específicos del sector. Estos frameworks incorporan conocimiento experto sobre patrones de amenazas y vulnerabilidades comunes, mejorando la eficacia del proceso de automatización. La integración de métricas específicas de seguridad en el

proceso de optimización garantiza que los modelos resultantes sean efectivos en la detección y mitigación de amenazas.

3.11.3 Inteligencia Artificial Explicable (XAI)

La incorporación de técnicas de XAI en sistemas de RAG y *Fine-Tuning* representa un avance crucial para la adopción generalizada de estas tecnologías en entornos críticos. Los modelos explicables proporcionan justificaciones claras para sus decisiones, permitiendo a los profesionales de seguridad comprender y validar las recomendaciones generadas. Esta transparencia resulta especialmente importante en sectores regulados donde la trazabilidad de las decisiones es un requisito legal.

Las técnicas de XAI aplicadas a la ciberseguridad facilitan la identificación de patrones de ataque y la comprensión de las vulnerabilidades detectadas. La capacidad de explicar el razonamiento detrás de las alertas de seguridad mejora la eficiencia en la respuesta a incidentes y reduce los falsos positivos. Además, la explicabilidad fortalece la confianza en los sistemas automatizados de seguridad, facilitando su adopción en organizaciones conservadoras.

3.11.4 Aprendizaje federado y privacidad

El aprendizaje federado emerge como una solución innovadora para abordar las preocupaciones de privacidad en la implementación de RAG y *Fine-Tuning*. Esta técnica permite el entrenamiento colaborativo de modelos sin centralizar datos sensibles, manteniendo la información crítica dentro de los perímetros de seguridad de cada organización. La implementación de protocolos de aprendizaje federado facilita la creación de modelos robustos mientras se cumple con regulaciones estrictas de protección de datos.

La integración de técnicas de privacidad diferencial en el aprendizaje federado proporciona garantías matemáticas sobre la protección de la información sensible. Estos mecanismos permiten el desarrollo de modelos efectivos mientras se mantiene un alto nivel de privacidad, facilitando la colaboración entre organizaciones en la lucha contra amenazas comunes. La combinación de aprendizaje federado con técnicas de agregación segura permite la creación de modelos más robustos sin comprometer la confidencialidad de los datos.

3.11.5 Convergencia tecnológica y estándares

El desarrollo de estándares y protocolos para la integración de RAG y *Fine-Tuning* en sistemas de ciberseguridad representa un paso crucial hacia la madurez del campo. La estandarización facilita la interoperabilidad entre diferentes implementaciones y permite la creación de ecosistemas de seguridad más cohesivos. El establecimiento de métricas comunes para evaluar la efectividad de estas tecnologías mejora la comparabilidad y facilita la selección de soluciones apropiadas.

La convergencia de estas tecnologías con otros avances emergentes, como la computación cuántica y blockchain, abre nuevas posibilidades para la seguridad informática. La integración de capacidades cuánticas podría mejorar significativamente la detección de patrones complejos, mientras que blockchain podría proporcionar un registro inmutable de eventos de seguridad y actualizaciones de modelos. Esta convergencia tecnológica promete crear sistemas de seguridad más resilientes y adaptables a las amenazas futuras.

3.12 Conclusión de la discusión

La revisión de los hallazgos en el uso de *Retrieval-Augmented Generation* (RAG), el *Fine-Tuning* y los métodos tradicionales dentro del ámbito de la ciberseguridad demuestra que

cada técnica ofrece ventajas y limitaciones particulares que deben analizarse a la luz de las necesidades específicas de cada contexto. Por un lado, RAG brinda la flexibilidad y la capacidad de respuesta necesarias para escenarios dinámicos, sobre todo cuando se cuenta con fuentes externas de datos confiables y actualizadas. El *Fine-Tuning*, por su parte, ha demostrado un alto rendimiento en dominios específicos y bien definidos, como el análisis de código fuente o la detección de amenazas en sectores críticos. Mientras tanto, los métodos tradicionales aún conservan su utilidad en entornos de menor complejidad, donde los recursos tecnológicos y presupuestarios son limitados.

Aun cuando estas técnicas han mostrado resultados alentadores, existen desafíos técnicos y de implementación que deben ser abordados. Por ejemplo, la dependencia de RAG en fuentes externas implica la necesidad de garantizar la calidad y la actualidad de estas bases de datos, mientras que sus altos costos computacionales pueden suponer una barrera de entrada para algunas organizaciones. El *Fine-Tuning*, por su parte, requiere datos cuidadosamente etiquetados y una infraestructura de entrenamiento sólida, lo cual puede dificultar su adopción en sectores con escasa documentación o recursos limitados. En el caso de los métodos tradicionales, su principal debilidad radica en la falta de adaptabilidad y la necesidad de actualizaciones manuales continuas en entornos altamente cambiantes.

Mirando hacia el futuro, se vislumbra un enorme potencial en la convergencia de RAG y *Fine-Tuning*, así como en la implementación de técnicas emergentes que fortalezcan estos enfoques. La creación de sistemas híbridos adaptativos podría combinar la inmediatez y contextualización de RAG con la precisión y la capacidad de personalización del *Fine-Tuning*, permitiendo una mayor eficacia y resiliencia frente a nuevas ciberamenazas. Asimismo, la

incorporación de sistemas explicables (XAI) y el aprendizaje federado ampliarían la transparencia y salvaguardarían la privacidad, requisitos fundamentales en sectores regulados y escenarios donde los datos son muy sensibles.

Por otra parte, la automatización del *Fine-Tuning* mediante plataformas AutoML posibilitaría que las organizaciones con menos recursos tecnológicos o experiencia en ciencia de datos puedan acceder a soluciones avanzadas de IA, reduciendo sustancialmente las barreras de adopción. Esta democratización de tecnologías contribuye no solo a potenciar la innovación en ciberseguridad, sino también a responder con rapidez a amenazas emergentes que ponen en jaque a todo tipo de organizaciones, desde pequeñas empresas hasta grandes corporaciones.

La implementación efectiva de RAG y *Fine-Tuning*, en conjunto o de manera individual, dependerá de una evaluación rigurosa de los requerimientos operativos, la disponibilidad de datos de calidad y la capacidad de inversión en infraestructura. El panorama actual indica que la combinación de estas técnicas, junto con metodologías que garanticen la interpretabilidad y la protección de la privacidad, puede redefinir significativamente la forma en que se gestionan las defensas cibernéticas. Aún queda camino por recorrer, pero los progresos alcanzados evidencian un futuro prometedor, en el que las soluciones híbridas, la automatización y la adopción de estándares comunes refuercen la capacidad de las organizaciones para responder de manera eficaz y oportuna a las constantes amenazas digitales.

Conclusiones

En primer lugar, se evidencia que las técnicas basadas en Modelos de Lenguaje de Gran Escala (LLMs), concretamente *Retrieval-Augmented Generation* (RAG) y *Fine-Tuning*, superan a los métodos de detección tradicionales en términos de precisión y reducción de falsos positivos. Esta superioridad se fundamenta en la habilidad de manejar volúmenes extensos de datos y patrones complejos, propiciando un análisis más completo de las amenazas emergentes. No obstante, se aprecian diferencias notables en cuanto a requerimientos y entornos de aplicación, lo que sugiere que la selección de la técnica depende en gran medida de las características específicas de cada organización. En este sentido, la idoneidad de cada enfoque se ve mediada por factores como la disponibilidad de datos, la infraestructura computacional y el nivel de expertise en inteligencia artificial.

En segundo lugar, RAG destaca por su adaptabilidad a escenarios altamente dinámicos, ya que integra datos externos y actualizados de manera continua, mejorando la respuesta ante vulnerabilidades emergentes. Este atributo resulta especialmente útil en infraestructuras críticas que exigen inmediatez en la identificación de amenazas, como el sector energético o el financiero. Sin embargo, dicha ventaja supone también retos operativos, especialmente en la gestión de bases de datos confiables, la optimización de recursos de cómputo y la minimización de la latencia en la recuperación de la información.

Por otro lado, el *Fine-Tuning* brinda oportunidades de especialización en dominios concretos, tales como la detección de fallos de seguridad en el código fuente o la identificación de brechas en entornos médicos. En estas situaciones, el ajuste fino de modelos previamente entrenados incrementa la exactitud y la pertinencia de la detección, aunque exige conjuntos de datos cuidadosamente etiquetados y una infraestructura de entrenamiento robusta. Además, el

riesgo de sobreajuste y la inversión de tiempo para reentrenar el modelo constituyen barreras que pueden afectar la aplicabilidad inmediata en entornos de cambio constante.

Asimismo, se constata que la combinación estratégica de RAG y *Fine-Tuning* abre un panorama prometedor para la ciberseguridad, al reunir la fuerza de la integración de información externa con la adaptabilidad puntual de los modelos ajustados. Este enfoque híbrido puede proporcionar soluciones escalables y precisas, al tiempo que potencia la capacidad de respuesta ante incidentes complejos o de rápida evolución. Sin embargo, su éxito depende de la existencia de marcos metodológicos y estándares claros, así como de la disponibilidad de recursos técnicos y humanos que respalden la adopción de estas metodologías avanzadas.

Además, la incorporación de Inteligencia Artificial Explicable (XAI) y el aprendizaje federado se erigen como tendencias clave para solventar las limitaciones en cuanto a interpretabilidad, privacidad de datos y escalabilidad. Estas aproximaciones permiten transparentar las decisiones del modelo y reforzar la protección de la información en sectores donde las normativas de confidencialidad son muy estrictas, ampliando así las posibilidades de adopción de sistemas basados en IA. Su implementación, no obstante, demanda lineamientos claros y consensuados que faciliten la integración de la explicabilidad y la seguridad en el corazón de las soluciones tecnológicas.

En síntesis, el panorama actual muestra a RAG y *Fine-Tuning* como herramientas valiosas para la defensa contra amenazas cibernéticas, pero su implementación efectiva exige un análisis cuidadoso de las necesidades organizativas, la infraestructura disponible y la calidad de los datos. La adopción exitosa de estos enfoques se verá reforzada por la convergencia de metodologías emergentes, la colaboración interdisciplinaria y la búsqueda continua de mejoras

en materia de optimización y transparencia de la inteligencia artificial aplicada a la ciberseguridad.

Recomendaciones

En primer lugar, se sugiere robustecer la calidad y la gobernanza de datos, de modo que se cuente con fuentes externas confiables para RAG y conjuntos de datos debidamente etiquetados en el caso de *Fine-Tuning*. Este paso implica el establecimiento de procesos de auditoría y actualización frecuentes, con el fin de minimizar la obsolescencia de la información y garantizar la vigencia de los modelos entrenados. Asimismo, resulta aconsejable invertir en herramientas de etiquetado semiautomático que agilicen la clasificación precisa de las muestras y reduzcan el impacto de posibles sesgos en la detección de vulnerabilidades.

En segundo lugar, conviene evaluar de manera detallada la infraestructura tecnológica requerida para la implementación de RAG y *Fine-Tuning*. Aquellas organizaciones con un alto volumen de datos y restricciones de latencia deben optar por soluciones de cómputo escalables, ya sea on-premise o en la nube, acompañadas de planes de contingencia que aseguren la continuidad del servicio. Adicionalmente, la exploración de técnicas como la compresión de modelos o la adopción de arquitecturas más ligeras puede contribuir a reducir los costos de operación sin sacrificar rendimiento en entornos productivos.

En tercer término, se recomienda diseñar estrategias de adopción híbridas que combinen RAG y *Fine-Tuning* en escenarios de elevada complejidad. Con este enfoque, el procesamiento de información en tiempo real (RAG) se integra con la fineza de modelos entrenados para contextos específicos (*Fine-Tuning*), potenciando la capacidad de respuesta ante incidentes novedosos. Para ello, resulta esencial establecer marcos de prueba piloto o laboratorios de innovación que validen la viabilidad técnica y económica, así como la eficacia de la detección en situaciones reales.

Por otra parte, se considera imperativo incorporar criterios de explicabilidad y privacidad en los proyectos de ciberseguridad basados en IA. El uso de técnicas de XAI permitirá a los equipos de seguridad entender los factores que impulsan la generación de alertas y priorizarlas de manera adecuada, lo cual redundará en una mejor toma de decisiones. De igual forma, la integración de aprendizaje federado es clave para evitar la exposición de datos sensibles, especialmente en sectores con estrictas regulaciones, favoreciendo la confianza y la colaboración entre distintas entidades.

También resulta fundamental promover la formación continua de profesionales de ciberseguridad en metodologías de IA avanzada y en la interpretación de modelos complejos. Ello abarca el fortalecimiento de programas académicos y la participación en foros de divulgación científica, propiciando la actualización constante y la conformación de equipos multidisciplinarios. Este factor humano cobra relevancia al momento de gestionar adecuadamente los modelos, su mantenimiento y la respuesta ante emergencias.

Finalmente, se recomienda la elaboración de estándares y lineamientos que orienten la adopción de RAG y *Fine-Tuning* a nivel institucional y gubernamental. Estas directrices deberían incluir protocolos de evaluación de rendimiento, recomendaciones de buenas prácticas y mecanismos de auditoría que fomenten la transparencia, la interoperabilidad y la capacidad de respuesta. Con ello, las organizaciones podrán aprovechar de manera más efectiva las ventajas competitivas que ofrece la inteligencia artificial en la protección y salvaguarda de activos digitales y de la información crítica.

Referencias bibliográficas

- Abdullahi, M., Baashar, Y., Alhussian, H., Alwadain, A., Aziz, N., Capretz, L. F., & Abdulkadir, S. J. (2022a). Detecting Cybersecurity Attacks in Internet of Things Using Artificial Intelligence Methods: A Systematic Literature Review. *Electronics*, 11(2). <https://doi.org/10.3390/ELECTRONICS11020198>
- Al-Hawawreh, M., Aljuhani, A., & Jararweh, Y. (2023). Chatgpt for cybersecurity: practical applications, challenges, and future directions. *Cluster Computing*, 26(6), 3421–3436. <https://doi.org/10.1007/S10586-023-04124-5>
- Ali, T. (2024b). Next-generation intrusion detection systems with LLMs : real-time anomaly detection, explainable AI, and adaptive data generation. *Laturi.Oulu.Fi*. <https://oulurepo.oulu.fi/handle/10024/51067>
- Bhatt, M., Chennabasappa, S., Li, Y., Nikolaidis, C., Song, D., Wan, S., Ahmad, F., Aschermann, C., Chen, Y., Kapil, D., Molnar, D., Whitman, S., & Saxe, J. (2024b). *CyberSecEval 2: A Wide-Ranging Cybersecurity Evaluation Suite for Large Language Models*. <https://arxiv.org/abs/2404.13161v1>
- Campbell, M., & Jovanovic, M. (2024). Disinfecting AI: Mitigating Generative AI's Top Risks. *Computer*, 57(5), 111–116. <https://doi.org/10.1109/MC.2024.3374433>
- Cao, Y., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., Li, S., & Sun, L. (2018). A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT; A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. *J. ACM*, 37(111).
- Capodiecì, N., Sanchez-Adames, C., Harris, J., & Tatar, U. (2024). The Impact of Generative AI and LLMs on the Cybersecurity Profession. *2024 Systems and Information Engineering Design Symposium, SIEDS 2024*, 448–453. <https://doi.org/10.1109/SIEDS61124.2024.10534674>
- Chakraborty, A., Biswas, A., & Khan, A. K. (2023). Artificial Intelligence for Cybersecurity: Threats, Attacks and Mitigation. *Intelligent Systems Reference Library*, 231, 3–25. https://doi.org/10.1007/978-3-031-12419-8_1
- Charfeddine, M., Kammoun, H. M., Hamdaoui, B., & Guizani, M. (2024). ChatGPT's Security Risks and Benefits: Offensive and Defensive Use-Cases, Mitigation Measures, and Future Implications. *IEEE Access*, 12, 30263–30310. <https://doi.org/10.1109/ACCESS.2024.3367792>
- Chen, J., Hu, S., Zheng, H., Xing, C., & Zhang, G. (2023). GAIL-PT: an intelligent penetration testing framework with generative adversarial imitation learning. *Comput. Secur.*, 126. <https://doi.org/10.1016/j.cose.2022.103055>
- Deshpande, A. S., & Gupta, S. (2023). GenAI in the Cyber Kill Chain: A Comprehensive Review of Risks, Threat Operative Strategies and Adaptive Defense Approaches. *3rd IEEE International Conference on ICT in Business Industry and Government, ICTBIG 2023*. <https://doi.org/10.1109/ICTBIG59752.2023.10456106>

- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). Opinion Paper: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642. <https://doi.org/10.1016/J.IJINFOMGT.2023.102642>
- Ferrag, M. A., Debbah, M., & Al-Hawawreh, M. (2023). Generative AI for Cyber Threat-Hunting in 6G-enabled IoT Networks. *Proceedings - 23rd IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing Workshops, CCGridW 2023*, 16–25. <https://doi.org/10.1109/CCGridW59191.2023.00018>
- Ferrara, E. (2024). GenAI against humanity: nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, 7(1), 549–569. <https://doi.org/10.1007/S42001-024-00250-1/TABLES/5>
- Gadyatskaya, O., & Papuc, D. (2023). ChatGPT Knows Your Attacks: Synthesizing Attack Trees Using LLMs. *Communications in Computer and Information Science*, 1942 CCIS, 245–260. https://doi.org/10.1007/978-981-99-7969-1_18
- Gill, S. S., & Kaur, R. (2023). ChatGPT: Vision and challenges. *Internet of Things and Cyber-Physical Systems*, 3, 262–271. <https://doi.org/10.1016/J.IOTCPS.2023.05.004>
- Guo, D., Chen, H., Wu, R., & Wang, Y. (2023). AIGC challenges and opportunities related to public safety: A case study of ChatGPT. *Journal of Safety Science and Resilience*, 4(4), 329–339. <https://doi.org/10.1016/J.JNLSSR.2023.08.001>
- Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. *IEEE Access*, 11, 80218–80245. <https://doi.org/10.1109/access.2023.3300381>
- Hilario, E., Azam, S., Sundaram, J., Imran Mohammed, K., & Shanmugam, B. (2024). Generative AI for pentesting: the good, the bad, the ugly. *International Journal of Information Security*, 23(3), 2075–2097. <https://doi.org/10.1007/S10207-024-00835-X/FIGURES/2>
- Hu, C., & Chen, J. (2023). A Dimensional Perspective Analysis on the Cybersecurity Risks and Opportunities of ChatGPT-Like Information Systems. *Proceedings - 2023 International Conference on Networking and Network Applications, NaNA 2023*, 324–331. <https://doi.org/10.1109/NANA60121.2023.00061>
- Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., & Neubig, G. (2023). Active Retrieval Augmented Generation. *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 7969–7992. <https://doi.org/10.18653/V1/2023.EMNLP-MAIN.495>
- Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D’Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B.,

- ... Zhao, S. (2021). Advances and Open Problems in Federated Learning. *Foundations and Trends® in Machine Learning*, 14(1–2), 1–210. <https://doi.org/10.1561/22000000083>
- Katsadourous, E., Patrikakis, C. Z., & Hurlburt, G. (2023). Can Large Language Models Better Predict Software Vulnerability? *IT Professional*, 25(3), 4–8. <https://doi.org/10.1109/MITP.2023.3284628>
- Kolosnjaji, B., Zarras, A., Webster, G., & Eckert, C. (2016). Deep learning for classification of malware system call sequences. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9992 LNAI, 137–149. https://doi.org/10.1007/978-3-319-50127-7_11
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems, 2020-December*. <https://arxiv.org/abs/2005.11401v4>
- Motlagh, F. N., Hajizadeh, M., Majd, M., Najafi, P., Cheng, F., & Meinel, C. (2024a). Large Language Models in Cybersecurity: State-of-the-Art. *Proceedings of ACM Conference (Conference '17)*, 1. <https://arxiv.org/abs/2402.00891v1>
- Mozes, M., He, X., Kleinberg, B., & Griffin, L. D. (2023). *Use of LLMs for Illicit Purposes: Threats, Prevention Measures, and Vulnerabilities*. <http://arxiv.org/abs/2308.12833>
- Mulder, V., Mermoud, A., & Lenders, V. (n.d.). *Andrei Kucharavy · Octave Plancherel · Large Language Models in Cybersecurity Threats, Exposure and Mitigation*.
- Nah, F. F.-H., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, 25(3), 277–304. <https://doi.org/10.1080/15228053.2023.2233814>
- Okey, O. D., Udo, E. U., Rosa, R. L., Rodríguez, D. Z., & Kleinschmidt, J. H. (2023). Investigating ChatGPT and cybersecurity: A perspective on topic modeling and sentiment analysis. *Computers & Security*, 135, 103476. <https://doi.org/10.1016/J.COSE.2023.103476>
- Pantserev, K. A. (2020). The malicious use of ai-based deepfake technology as the new threat to psychological security and political stability. *Advanced Sciences and Technologies for Security Applications*, 37–55. https://doi.org/10.1007/978-3-030-35746-7_3
- Pasupuleti, R., Vadapalli, R., & Mader, C. (2023). Cyber Security Issues and Challenges Related to Generative AI and ChatGPT. *Proceedings - 2023 10th International Conference on Social Networks Analysis, Management and Security, SNAMS 2023*. <https://doi.org/10.1109/SNAMS60348.2023.10375472>
- Rani, V., Kumar, M., Mittal, A., & Kumar, K. (2022). Artificial Intelligence for Cybersecurity: Recent Advancements, Challenges and Opportunities. *Studies in Computational Intelligence*, 1030, 73–88. https://doi.org/10.1007/978-3-030-96737-6_4

- Saddi, V. R., Gopal, S. K., Mohammed, A. S., Dhanasekaran, S., & Naruka, M. S. (2024). Examine the Role of Generative AI in Enhancing Threat Intelligence and Cyber Security Measures. *2024 2nd International Conference on Disruptive Technologies, ICDT 2024*, 537–542. <https://doi.org/10.1109/ICDT61202.2024.10489766>
- Sai, S., Yashvardhan, U., Chamola, V., & Sikdar, B. (2024). Generative AI for Cyber Security: Analyzing the Potential of ChatGPT, DALL-E, and Other Models for Enhancing the Security Space. *IEEE Access*, 12, 53497–53516. <https://doi.org/10.1109/ACCESS.2024.3385107>
- Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions. *SN Computer Science*, 2(3). <https://doi.org/10.1007/S42979-021-00557-0>
- Scanlon, M., Breiting, F., Hargreaves, C., Hilgert, J. N., & Sheppard, J. (2023). ChatGPT for digital forensic investigation: The good, the bad, and the unknown. *Forensic Science International: Digital Investigation*, 46, 301609. <https://doi.org/10.1016/J.FSIDI.2023.301609>
- Sewak, M., Emani, V., & Naresh, A. (2023a). *CRUSH: Cybersecurity Research using Universal LLMs and Semantic Hypernetworks*. <https://www.linkedin.com/in/mohitsewak/>
- Singh, J. P., & Agrawal, S. (2024). Automating Threat Intelligence Analysis with Retrieval Augmented Generation (RAG) for Enhanced Cybersecurity Posture. *Article in International Journal of Science and Research*. <https://doi.org/10.21275/SR24502103758>
- Sophos 2024 Threat Report: Cybercrime on Main Street. (n.d.).
- Verma, S. (2023). *Redefining Network Security: A Comparative Study Of Traditional And AI-Driven Approaches*. 11, 2320–2882. www.ijcrt.org
- Xu, H., Wang, S., Li, N., Wang, K., Zhao, Y., Chen, K., Yu, T., Liu, Y., & Wang, H. (2024a). *Large Language Models for Cyber Security: A Systematic Literature Review*. 1. <https://arxiv.org/abs/2405.04760v2>
- Yamin, M. M., Hashmi, E., Ullah, M., & Katt, B. (2024a). *Applications of LLMs for Generating Cyber Security Exercise Scenarios*. <https://doi.org/10.21203/RS.3.RS-3970015/V1>
- Zhang, J., Bu, H., Wen, H., Chen, Y., Li, L., & Zhu, H. (2024a). *When LLMs Meet Cybersecurity: A Systematic Literature Review*. <https://arxiv.org/abs/2405.03644v1>
- Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2023). Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1), 237–291. https://doi.org/10.1162/coli_a_00502