Universidad Internacional de La Rioja

Escuela Superior de Ingeniería y Tecnología

Máster Universitario en Ciberseguridad

Evaluación del Impacto de los LLM en la Ciberseguridad: Mejoras y Riesgos Potenciales.

Trabajo fin de estudio presentado por:	SEBASTIAN CARDENAS
Tipo de trabajo:	Tipo 1. Piloto experimental
Director/a:	MANUEL GARCIA FLORES
Fecha:	23/10/2024

Resumen

Este piloto experimental examina la confiabilidad y eficacia de los Modelos de Lenguaje de Gran Escala (LLMs) en la configuración de firewalls de próxima generación, ofreciendo una visión comparativa con los métodos convencionales en ciberseguridad. Para ello, se identificaron lineamientos de referencia a partir de organizaciones como el NIST y el CIS, y se seleccionaron cuatro LLMs (ChatGPT, Gemini, LLaMA 3 y Claude) con el propósito de generar configuraciones de seguridad en escenarios reales. Dichas configuraciones se evaluaron frente a estándares y guías oficiales de dos fabricantes líderes (Fortinet y Check Point). El análisis incluyó políticas de control de acceso, aplicaciones, prevención de intrusiones (IPS) y autenticación basada en identidad de usuario. Los resultados indican que, si bien los LLMs producen configuraciones iniciales en menor tiempo y con un grado razonable de adecuación, persisten carencias en parámetros avanzados y en la detección de amenazas cifradas. Estas omisiones pueden incrementar la exposición a riesgos, por lo que la validación experta sigue siendo necesaria. Aun así, se concluye que los LLMs constituyen un recurso de apoyo valioso para profesionales con menor experiencia, al permitir una curva de aprendizaje acelerada y facilitar la implementación de políticas de firewall. Se propone, finalmente, profundizar en la alineación de estos modelos con prácticas de seguridad más exigentes y explorar métodos automatizados de verificación y corrección de configuraciones.

Palabras clave: Inteligencia artificial, Modelos de lenguaje, Ciberseguridad, Firewalls de nueva generación, Configuración segura.

Abstract

This study examines the reliability and effectiveness of Large Language Models (LLMs) in configuring next-generation firewalls, offering a comparative perspective with conventional methods in cybersecurity. To this end, reference guidelines from organizations such as NIST and CIS were identified, and four LLMs (ChatGPT, Gemini, LLaMA 3, and Claude) were selected to generate security configurations in real-world scenarios. These configurations were assessed against standards and official guides from two leading manufacturers (Fortinet and Check Point). The analysis included policies for access control, application control, intrusion prevention (IPS), and user identity-based authentication. The results indicate that, while LLMs can produce initial configurations in less time with a reasonable degree of adequacy, shortcomings persist in advanced parameters and in detecting encrypted threats. These omissions may increase exposure to risks, highlighting the continued need for expert validation. Nevertheless, it is concluded that LLMs represent a valuable resource to support less-experienced professionals by accelerating the learning curve and facilitating firewall policy implementation. Finally, this work proposes further research to align these models with more demanding security practices and to explore automated methods for verification and correction of configurations.

Keywords: Large Language Models, Cybersecurity, Next-Generation Firewalls, Best Practices, Configuration Automation.

Índice de contenidos

Índice de figuras

Índice de tablas

1. Introducción

La presente investigación se enmarca en la coyuntura entre innovación y riesgo que representa la integración de soluciones de Inteligencia Artificial (IA) en el ámbito de la ciberseguridad, con especial atención a los Modelos de Lenguaje de Gran Escala (LLMs, por sus siglas en inglés).

La creciente complejidad de las amenazas cibernéticas y la evolución constante de las tecnologías de defensa han generado un escenario en el que las organizaciones necesitan soluciones cada vez más sofisticadas para proteger sus activos digitales. En este contexto, la Inteligencia Artificial (IA) y, más recientemente, los LLMs, han cobrado un protagonismo notable. Tradicionalmente, las técnicas de machine learning y deep learning se han implementado en la detección de amenazas y la protección ante ataques de día cero; sin embargo, los LLMs añaden una capa novedosa de interacción con el lenguaje, lo que abre la puerta a formas más automatizadas y flexibles de gestionar configuraciones y políticas de seguridad.

El uso de estas herramientas se ha expandido con rapidez en entornos educativos y profesionales, incluyendo el ámbito de la ciberseguridad. La posibilidad de obtener respuestas detalladas y aparentemente precisas con un mínimo de instrucciones escritas ha resultado atractiva para profesionales de todos los niveles de experiencia, en especial para aquellos con menos formación técnica. No obstante, existen preocupaciones legítimas sobre la exactitud de las respuestas proporcionadas por los LLMs y sobre los posibles riesgos de depender en exceso de sistemas que, si bien pueden ser muy completos en su generación de texto, no necesariamente garantizan conformidad con buenas prácticas o estándares de seguridad.

En este sentido, cualquier fallo en la configuración de políticas de seguridad puede derivar en vulnerabilidades críticas, especialmente en infraestructuras que administran información sensible o sistemas estratégicos.

Por ello, la presente investigación busca analizar, de manera comparativa y sistemática, la capacidad de los LLMs para generar configuraciones de firewall en escenarios representativos de la administración de la ciberseguridad. Se busca verificar la capacidad de dichas herramientas para generar configuraciones alineadas con estándares y buenas prácticas, evaluar su eficacia y determinar los nuevos riesgos que surgen de su uso.

1.1. Motivación

En los últimos años el papel de las herramientas de inteligencia artificial dentro del área de la ciberseguridad se ha diversificado, a las ya conocidas aplicaciones de deep learning y machine learning para la detección de amenazas y protección contra ataques de día cero, se unen ahora las aplicaciones de Large Language Models - LLM en ciberseguridad.

La popularidad y relevancia de los LLMs ha crecido de manera exponencial dentro de los entornos educativos y profesionales de manera transversal incluyendo por supuesto a la ciberseguridad, esto se evidencia en el creciente uso de herramientas LLM entre profesionales de la ciberseguridad, con mayor impacto en los grupos con menor experiencia.

Entre las principales causas de esta tendencia se encuentran la accesibilidad y facilidad de uso que ofrecen los LLMs. Estas herramientas permiten obtener soluciones automáticas a problemas complejos, sin que el usuario deba poseer conocimientos técnicos profundos. Sin embargo, las respuestas rápidas y detalladas que ofrece las LLM pueden no ser lo suficientemente precisas o pueden omitir aspectos fundamentales de la ciberseguridad (Zhang et al., 2023). La relevancia de este problema radica en que la ciberseguridad es un campo donde cualquier error en la operación o gestión de sistemas de seguridad puede exponer a una organización a vulnerabilidades significativas.

Por lo tanto, esta investigación busca evaluar el desempeño de los LLMs en comparación con los métodos convencionales en gestión de ciberseguridad, identificando no solo su eficacia, sino también los posibles nuevos riesgos asociados con su uso indebido (He et al., 2023; Lin & Wang, 2023). La creciente dependencia de estas herramientas subraya la importancia de abordar tanto sus capacidades como las limitaciones en entornos de seguridad críticos (Sadiku et al., 2020).

1.2. Planteamiento del problema

Dentro del sector de la ciberseguridad el uso creciente de herramientas LLM debido a su capacidad para automatizar tareas complejas. Sin embargo, esta popularidad presenta un reto ¿son confiables y seguras las soluciones generadas por LLMs en comparación con los enfoques tradicionales empleados por expertos en ciberseguridad?

Con las LLM existe el riesgo de que las configuraciones resultantes no sean precisas o adecuadas para entornos críticos. Además, los cibercriminales también están aprovechando estas herramientas para desarrollar ataques más sofisticados. Por tanto, es fundamental evaluar el impacto real de los LLMs en la ciberseguridad para determinar si pueden utilizarse de manera segura y eficaz.

Este trabajo propone realizar un análisis comparativo entre las soluciones generadas por LLMs en diferentes escenarios de la gestión de ciberseguridad. El objetivo es verificar si los resultados obtenidos por los LLMs pueden cumplir con los estándares de seguridad requeridos y proporcionar un marco para su implementación responsable en entornos de ciberseguridad.

1.3. Estructura del trabajo

El presente estudio se organiza en cinco capítulos que abordan de manera integral la evaluación de los LLMs en la configuración de soluciones de ciberseguridad.

En el estado del arte, se realiza una revisión exhaustiva de la literatura científica relevante, analizando investigaciones sobre el impacto de los LLMs en la ciberseguridad. Se identifican tanto las mejoras aportadas como los riesgos potenciales asociados con su implementación, proporcionando un contexto fundamental y estableciendo las bases teóricas para la investigación.

En el siguiente capitulo se establece el objetivo general y los objetivos específicos de la investigación, detallando la metodología adoptada para alcanzarlos. Se describe la selección de los LLMs, la definición de los escenarios de configuración, la incorporación de fabricantes específicos y la estrategia para comparar los resultados con estándares y buenas prácticas reconocidas.

En el capítulo cuatro, se presenta el desarrollo detallado del experimento piloto realizado. Se describe el proceso de interacción con los LLMs, la obtención y evaluación de las configuraciones generadas, y la comparación con los estándares de seguridad establecidos. Se analizan los resultados obtenidos, identificando posibles riesgos y oportunidades asociados con el uso de LLMs en entornos de ciberseguridad.

Finalmente, en conclusiones y trabajos futuros, se ofrecen las conclusiones de la investigación y se proponen líneas de trabajo futuro. Se resume el alcance y relevancia de la contribución realizada, discutiendo el grado de cumplimiento de los objetivos planteados y las implicaciones prácticas de los hallazgos obtenidos.

2. Estado del arte

La rápida evolución del panorama digital ha llevado a un incremento exponencial en la cantidad y sofisticación de las amenazas cibernéticas. En este contexto, la Inteligencia Artificial (IA), y específicamente los Modelos de Lenguaje de Gran Escala (LLMs, por sus siglas en inglés), han emergido como herramientas cruciales en el ámbito de la ciberseguridad. Estos modelos ofrecen capacidades avanzadas para la detección, prevención y respuesta a amenazas, pero también introducen nuevos riesgos y desafíos. Este estado del arte tiene como objetivo analizar críticamente las investigaciones existentes sobre el impacto de los LLMs en la ciberseguridad, identificando tanto las mejoras aportadas como los riesgos potenciales asociados con su implementación.

Para abordar este análisis, se ha llevado a cabo una revisión exhaustiva de la literatura científica relevante, siguiendo una estrategia de búsqueda sistemática en bases de datos académicas como arXiv y revistas especializadas en ciberseguridad e inteligencia artificial. Se utilizaron palabras clave como "LLMs", "ciberseguridad", "inteligencia artificial", "amenazas cibernéticas" y "modelos de lenguaje". Se seleccionaron artículos que aportan perspectivas teóricas y empíricas sobre el uso de LLMs en ciberseguridad, asegurando la inclusión de estudios clave y la identificación de tendencias actuales en el campo.

Entre las investigaciones más relevantes, destaca el trabajo de Sadiku, Fagbohungbe y Musa (2020), quienes exploran la intersección entre la IA y la ciberseguridad. En su estudio, resaltan cómo herramientas como sistemas expertos, aprendizaje automático y procesamiento de lenguaje natural refuerzan las capacidades para salvaguardar redes e información crítica. La IA mejora significativamente las respuestas a incidentes en entornos dinámicos y de alta complejidad, aunque también presenta desafíos en términos de costos y barreras técnicas. Este análisis proporciona un contexto fundamental sobre cómo la IA ha sido instrumental en fortalecer las defensas cibernéticas y establece una base para comprender el papel específico de los LLMs.

De manera complementaria, Zhang et al. (2025) presentan un extenso análisis en el que revisan más de 300 artículos académicos, 25 modelos de lenguaje y 10 escenarios de aplicación en ciberseguridad. Su investigación busca dar respuesta a tres preguntas principales: la construcción de LLMs específicos para tareas de seguridad, las aplicaciones potenciales de dichos modelos en el ámbito defensivo y ofensivo y las futuras líneas de investigación. A través de distintas técnicas de ajuste, incluyendo entrenamientos continuos y enfoques ligeros como LoRA, demuestran que la incorporación de datos de ciberseguridad mejora la precisión en la detección de vulnerabilidades, la generación de código seguro y el análisis de malware. Sin embargo, enfatizan también limitaciones importantes, como la fiabilidad de las respuestas y la necesidad de regular los usos maliciosos de los LLMs. El estudio concluye que, si bien estos modelos pueden optimizar de manera notable procesos como la generación automatizada de inteligencia de amenazas o el hardening de software, persisten retos significativos en términos de interpretabilidad y regulación.

Por otro lado, Abdali et al. (2023) analizan las amenazas y vulnerabilidades asociadas con los LLMs, clasificándolas en categorías basadas en el modelo, tiempo de entrenamiento y tiempo de inferencia. Identifican riesgos como la fuga de información sensible, la memorización de datos de entrenamiento y la generación de código inseguro. Subrayan cómo estas debilidades pueden ser explotadas para comprometer la seguridad de sistemas y aplicaciones que dependen de estos modelos. Además, enfatizan la necesidad de estrategias de mitigación como el red teaming, la edición de modelos y el uso de marcas de agua en el texto generado por LLMs. Sin embargo, también destacan las limitaciones actuales de estas medidas y la importancia de una aproximación ética y responsable en el diseño y uso de LLMs.

En una línea similar a los hallazgos sobre la generación de código inseguro (Abdali et al., 2023), Sandoval et al. (2023) examinan cómo la asistencia de LLMs en programación puede introducir o exacerbar vulnerabilidades en el software. En su experimento, 58 desarrolladores, divididos en grupo control y grupo asistido con Codex LLM, debían implementar funciones en C. Si bien quienes contaban con la ayuda del LLM superaron más pruebas funcionales y compilaron con mayor éxito, el análisis de seguridad posterior reveló que los errores críticos provenían tanto del modelo como de la interacción humana con el código sugerido. Los autores sostienen que

las prácticas de revisión, la capacitación en seguridad y el perfeccionamiento del entrenamiento de los LLMs con datos de calidad son cruciales para mitigar los riesgos potenciales. Estos resultados subrayan la necesidad de combinar la eficiencia que ofrecen los asistentes de IA con métodos de aseguramiento y validación robustos que minimicen la posibilidad de errores explotables.

En un análisis crítico sobre el impacto de los modelos de IA generativa en la ciberseguridad y la privacidad, Gupta et al. (2023) resaltan cómo herramientas como ChatGPT pueden ser explotadas para facilitar ataques como ingeniería social, phishing y generación automatizada de malware. A través de casos específicos, demuestran la doble naturaleza de estas tecnologías, que pueden ser utilizadas tanto para mejorar las defensas cibernéticas como para potenciar las capacidades de los atacantes. Exploran también las vulnerabilidades inherentes a estas herramientas, incluyendo su susceptibilidad a ataques de inyección de comandos y métodos de manipulación como el jailbreaking. Este estudio subraya la necesidad de equilibrar el potencial transformador de los modelos generativos con medidas de seguridad y ética robustas.

Ji et al. (2023) presentan el conjunto de datos BEAVERTAILS, diseñado para alinear los LLMs con estándares de seguridad y valores humanos. Este dataset introduce una innovación significativa al separar las métricas de utilidad y seguridad en las interacciones de pregunta-respuesta, ofreciendo un enfoque granular para evaluar estos atributos críticos. Con más de 330,000 pares de datos y 360,000 comparaciones de preferencia humana, BEAVERTAILS destaca por su capacidad para mejorar modelos de lenguaje en tareas como moderación de contenido y aprendizaje por refuerzo basado en retroalimentación humana. Los autores demuestran mejoras significativas en la producción de respuestas útiles y seguras, aunque reconocen limitaciones en la representación demográfica de los datos y en la complejidad de algunas categorías de daño. Este trabajo no solo contribuye a la investigación en alineación de seguridad, sino que también enfatiza la necesidad de enfoques multidimensionales que equilibren utilidad, seguridad y ética en el despliegue de tecnologías de inteligencia artificial.

Por su parte, Zhou et al. (2023) investigan la tendencia de los LLMs a mostrar un sesgo de exceso de confianza en sus respuestas y su reticencia a expresar incertidumbre. A través de un enfoque experimental, los autores descubren que los modelos suelen evitar expresar incertidumbre, incluso cuando las respuestas son incorrectas, y que, al utilizar marcadores de confianza, frecuentemente recurren a expresiones de certeza excesiva. Esto resulta en tasas de error considerables y promueve una dependencia inapropiada de las respuestas generadas por IA. Identifican el aprendizaje por refuerzo con retroalimentación humana (RLHF) como un factor clave en esta propensión, influenciado por sesgos en los datos de entrenamiento y las preferencias humanas. El trabajo concluye con recomendaciones para mejorar la transparencia y la calibración de los LLMs, sugiriendo la inclusión proactiva de marcadores de incertidumbre y el diseño de interacciones que mitiguen la sobreconfianza de los modelos y el exceso de confianza de los usuarios en sus respuestas.

Si nos referimos a aplicaciones prácticas en seguridad, Deng et al. (2023) exploran las capacidades de los LLMs en el pentesting automatizado mediante la herramienta PENTESTGPT. Destacan cómo la comprensión contextual y generación de respuestas complejas por parte de los LLMs puede transformar prácticas tradicionalmente manuales y dependientes de expertos humanos. A través de un diseño innovador que incorpora módulos especializados de razonamiento, generación y análisis, PENTESTGPT propone una solución interactiva y eficiente para abordar tareas de prueba de penetración en escenarios reales y desafíos de tipo CTF. El estudio subraya que PENTESTGPT supera significativamente a los LLMs utilizados de manera directa, logrando tasas de finalización de tareas sustancialmente superiores y reduciendo la dependencia de conocimientos especializados. Sin embargo, también identifica desafíos inherentes, como la pérdida de contexto en tareas prolongadas y la generación de comandos imprecisos o herramientas inexistentes.

En cuanto a los ejercicios prácticos de seguridad, Tann et al. (2023) estudiaron el desempeño de varios LLMs (incluyendo ChatGPT, Bard y Bing) en la resolución de desafíos tipo Capture-The-Flag (CTF) y en preguntas de certificaciones profesionales de ciberseguridad (por ejemplo, CCNA y CCNP de Cisco). Su análisis reveló que, si bien los LLMs muestran un rendimiento destacable en preguntas fácticas y en ciertos retos de CTF (como análisis forense o criptografía

básica), tienen dificultades con desafíos que exigen razonamiento avanzado o la ejecución de pasos interactivos, como inyección de comandos y explotación web compleja. Asimismo, demostraron que los LLMs pueden ser sometidos a técnicas de "jailbreaking", lo que potencialmente facilita la generación de exploits maliciosos.

En el ámbito comercial, empresas como Check Point y SentinelOne han desarrollado soluciones basadas en LLMs que transforman las operaciones de equipos de ciberseguridad especializados. Infinity AI Copilot de Check Point se posiciona como una solución generativa que acelera hasta en un 90% las actividades de seguridad, permitiendo a los equipos de TI realizar actualizaciones de políticas, cazar amenazas y responder a incidentes de manera guiada por IA. Su diseño centrado en la privacidad, con una arquitectura privacy-by-design, garantiza que los datos de los usuarios no se compartan ni utilicen para entrenar futuros modelos, destacándose como una solución orientada tanto a la eficiencia como a la protección de la información sensible (Check Point, s.f.). Por su parte, Purple AI de SentinelOne introduce un analista de seguridad impulsado por IA que simplifica investigaciones complejas mediante consultas en lenguaje natural y generación automática de resúmenes. Esta herramienta integra análisis de amenazas y colaboración a través de cuadernos compartidos, permitiendo a los analistas reducir el tiempo medio de detección (MTTD) y acelerar la respuesta ante riesgos emergentes. Purple AI sobresale por su capacidad de traducir datos estructurados y no estructurados en perspectivas procesables, apoyándose en un marco de esquemas abiertos como OCSF para normalizar datos de múltiples fuentes (SentinelOne, s.f.).

Desde otro vector de la ciberseguridad, He et al. (2023) examinan el impacto de los LLMs en la seguridad de blockchain, destacando su capacidad para transformar estrategias de protección frente a amenazas cibernéticas en este dominio emergente. Presentan una revisión sistemática de la literatura existente sobre el uso de LLMs en tareas específicas como la auditoría de contratos inteligentes, la detección de anomalías en transacciones y el análisis dinámico. Este enfoque multidimensional resalta cómo los LLMs no solo optimizan la identificación de vulnerabilidades, sino que también ofrecen soluciones adaptativas y contextualmente enriquecidas que refuerzan la seguridad de los sistemas blockchain en un panorama tecnológico en constante evolución. Asimismo, abordan los desafíos éticos,

técnicos y de escalabilidad inherentes a la implementación de LLMs en este dominio, incluyendo preocupaciones sobre la privacidad de los datos, el consumo energético y las limitaciones en la generalización de los modelos. Los autores subrayan la necesidad de marcos regulatorios sólidos y de una colaboración interdisciplinaria para maximizar el potencial de estas herramientas, mitigando riesgos y ampliando su aplicabilidad.

El estudio de la literatura revela que los LLMs ofrecen mejoras significativas en la ciberseguridad, desde la detección avanzada de amenazas hasta la automatización de procesos complejos como el pentesting. Estos modelos aportan una capacidad sin precedentes para procesar y analizar grandes volúmenes de datos, identificar patrones anómalos y proporcionar respuestas adaptativas en tiempo real. Sin embargo, también introducen riesgos potenciales, incluyendo vulnerabilidades explotables por actores maliciosos, problemas de sesgo de exceso de confianza y desafíos éticos relacionados con la privacidad y el manejo de la incertidumbre.

Es evidente la necesidad de continuar investigando en áreas como la alineación ética y de seguridad de los LLMs, así como en el desarrollo de estrategias de mitigación efectivas. La implementación de conjuntos de datos como BEAVERTAILS y la adopción de prácticas de diseño centradas en la privacidad son pasos en la dirección correcta, pero se requiere un esfuerzo concertado que involucre a investigadores, profesionales de la industria y entidades regulatorias. La colaboración interdisciplinaria es esencial para abordar los desafíos identificados y promover el uso responsable y seguro de los LLMs en ciberseguridad.

Los vacíos identificados en la investigación actual, como las limitaciones en la generalización de los modelos y la necesidad de marcos regulatorios sólidos, representan oportunidades para futuras investigaciones. Es fundamental que los desarrollos tecnológicos vayan acompañados de consideraciones éticas y prácticas responsables para maximizar los beneficios y minimizar los riesgos asociados con la implementación de LLMs en ciberseguridad. Solo a través de un enfoque equilibrado y consciente se podrá aprovechar plenamente el potencial transformador de los LLMs en la protección de los entornos digitales.

3. Objetivos concretos y metodología de trabajo

3.1. Objetivo general

El objetivo general de esta investigación es verificar la confiabilidad y efectividad de los LLMs como asistentes en la gestión y configuración de herramientas de ciberseguridad. Se busca evaluar su capacidad para asistir en la correcta configuración de políticas de seguridad para firewall de nueva generación y sus distintas herramientas de seguridad. Además, se comparará su rendimiento con los estándares y buenas prácticas establecidos en el ámbito de la ciberseguridad, para determinar si los LLMs pueden ser una herramienta confiable para la gestión segura y eficiente de plataformas de seguridad.

3.2. Objetivos específicos

Evaluar cómo los LLMs pueden asistir a estudiantes y profesionales inexpertos en la configuración de herramientas de ciberseguridad.

Comparar los resultados generados por los LLMs con los estándares y buenas prácticas de seguridad vigentes, como los establecidos por organismos como NIST, ISO y CIS.

Identificar errores comunes o vulnerabilidades introducidos al utilizar LLMs para la configuración de políticas de seguridad.

Determinar la efectividad de los LLMs en la creación de configuraciones seguras en comparación con administradores humanos con poca experiencia.

Medir el impacto de los LLMs en términos de ahorro de tiempo y precisión, al comparar configuraciones asistidas por LLMs con configuraciones manuales.

Explorar el potencial de los LLMs para complementar el aprendizaje de los profesionales, acelerando su curva de aprendizaje y proporcionándoles soporte continuo en la toma de decisiones de configuración y gestión de seguridad.

3.3. Metodología del trabajo

El presente estudio se propone evaluar el impacto de los Modelos de Lenguaje de Gran Escala (LLMs) en la configuración de soluciones de ciberseguridad, específicamente en el contexto de los firewalls de nueva generación. Para alcanzar este objetivo, se ha diseñado una metodología estructurada en varias etapas clave, que incluyen la selección de los LLMs a utilizar, la definición de los escenarios de configuración, la incorporación de fabricantes específicos y la comparación de los resultados con estándares y buenas prácticas reconocidas.

En primer lugar, se procederá a la selección de los LLMs que formarán parte del estudio. Se han elegido cuatro modelos de vanguardia, todos en sus versiones de suscripción no gratuitas, para garantizar el acceso a sus capacidades completas. Los modelos seleccionados son:

- ChatGPT de OpenAI.
- Gemini de Google DeepMind.
- LLaMA 3 de Meta AI.
- Claude 3 de Anthropic.

La elección de estos modelos se basa en su relevancia en el campo de la inteligencia artificial, su amplia aceptación en la industria y su potencial para generar respuestas complejas y contextualmente relevantes en tareas técnicas.

A continuación, se definirán cinco escenarios de configuración de firewalls de nueva generación, diseñados para evaluar la capacidad de los LLMs en la generación de configuraciones de seguridad. Los escenarios se centrarán en los siguientes tipos de políticas:

- Políticas de Control de Acceso: Configuración de reglas que permiten o deniegan tráfico basado en direcciones IP, puertos y protocolos.
- Políticas de Control de Aplicaciones: Definición de reglas que gestionan el acceso a aplicaciones específicas, permitiendo un control granular del tráfico de aplicaciones.
- Políticas Basadas en Identidad del Usuario: Implementación de políticas que restringen o permiten el acceso según la identidad del usuario, integrando servicios de autenticación.
- Políticas de Prevención de Intrusiones (IPS): Configuración de sistemas que detectan y previenen actividades maliciosas o sospechosas en la red.

- Políticas Basadas en Geolocalización: Establecimiento de reglas que controlan el tráfico basado en la ubicación geográfica de origen o destino.

Estos escenarios reflejan situaciones comunes y relevantes en la administración de seguridad de redes, permitiendo evaluar la efectividad de los LLMs en diferentes aspectos críticos de la configuración de firewalls.

Cada uno de los escenarios se presentará a los LLMs en tres modalidades:

- De manera genérica: Se describirá el escenario sin especificar ningún fabricante, solicitando al modelo una configuración que aplique a un firewall de nueva generación en términos generales.
- Específicamente para el fabricante Fortinet: Se indicará que la configuración debe aplicarse en dispositivos Fortinet, esperando que el modelo genere instrucciones o comandos específicos para esta plataforma.
- Específicamente para el fabricante Check Point: Se solicitará la configuración adaptada a dispositivos Check Point, buscando evaluar la capacidad del modelo para ajustarse a diferentes entornos tecnológicos.

Para los escenarios específicos de Fortinet y Check Point, los LLMs que lo permitan recibirán como entrada el documento oficial de la guía de administración de las soluciones respectivas. Esta información adicional tiene como objetivo facilitar al modelo el acceso a detalles técnicos y sintaxis específicas de cada fabricante, permitiendo una generación de configuraciones más precisas y alineadas con las prácticas recomendadas por los proveedores.

Una vez obtenidas las configuraciones propuestas por los LLMs para cada escenario y fabricante, se procederá a su comparación con los estándares y buenas prácticas establecidos. Se utilizarán como referencia:

- CIS Benchmarks para Fortinet y Check Point: Estos benchmarks proporcionan guías detalladas de configuración segura específicas para cada fabricante, estableciendo parámetros y recomendaciones para asegurar que los dispositivos cumplen con niveles adecuados de seguridad.
- NIST SP 800-41 (Guidelines on Firewalls and Firewall Policy): Este documento del Instituto Nacional de Estándares y Tecnología ofrece directrices generales sobre la implementación y gestión de firewalls y políticas de seguridad asociadas.

La comparación se realizará mediante una matriz de análisis, que cruzará los resultados de los LLMs con los fabricantes y los estándares seleccionados. Esta matriz permitirá identificar:

- Conformidad: Grado en que las configuraciones propuestas cumplen con las recomendaciones y requisitos establecidos en los estándares y benchmarks.
- Errores u omisiones: Identificación de configuraciones incorrectas, faltantes de parámetros críticos o implementación de prácticas inseguras.
- Consistencia entre modelos: Análisis de la coherencia de las respuestas entre los diferentes LLMs, evaluando si existen patrones comunes o discrepancias significativas.
- Adaptación al fabricante: Capacidad de los LLMs para ajustar las configuraciones según las especificaciones y sintaxis de cada fabricante, utilizando la información proporcionada en las guías de administración.

El análisis de los resultados se centrará en determinar el impacto de los LLMs en la ciberseguridad, identificando tanto las mejoras potenciales como los riesgos asociados con su uso en procesos de configuración de soluciones de seguridad. Se explorará si los LLMs pueden ofrecer configuraciones que cumplan con los estándares de seguridad, si facilitan el proceso de configuración para usuarios inexpertos y si existen limitaciones o peligros en confiar en estas herramientas sin una supervisión adecuada.

Esta metodología permite abordar de manera sistemática los objetivos del estudio, proporcionando un marco claro para la evaluación de los LLMs en un contexto práctico y relevante. Al comparar las configuraciones generadas con estándares reconocidos y al considerar las particularidades de fabricantes específicos, se obtiene una perspectiva integral del desempeño de los modelos y de su aplicabilidad en entornos reales de ciberseguridad.

4. Desarrollo específico de la contribución

En este capítulo se presenta el desarrollo detallado del piloto realizado en esta investigación. Se describe de manera técnica el experimento llevado a cabo para evaluar el desempeño de los LLMs en la configuración de políticas de seguridad en firewalls de nueva generación. Este análisis se enfoca en verificar la confiabilidad y efectividad de los LLMs, y en identificar posibles riesgos asociados con su uso en entornos de ciberseguridad.

4.1. Tipo 1. Piloto experimental

El experimento se diseñó con el objetivo de evaluar la capacidad de los LLMs para generar configuraciones de seguridad que cumplan con los estándares y buenas prácticas establecidos por organismos reconocidos en el ámbito de la ciberseguridad, como el NIST y el CIS. Se buscó replicar escenarios reales en los que profesionales inexpertos podrían utilizar LLMs para configurar políticas de seguridad en dispositivos de firewall de nueva generación.

4.1.1. Descripción detallada del experimento

4.1.1.1. Selección de los Modelos de Lenguaje

Se seleccionaron cuatro LLMs de vanguardia, todos en sus versiones de suscripción por pago, para garantizar el acceso a las capacidades de su LLM más reciente:

ChatGPT o4 de OpenAI, cuenta con aproximadamente 175 mil millones de parámetros. Fue entrenado con una amplia variedad de datos textuales hasta mayo de 2024, lo que le permite comprender y generar texto coherente en múltiples idiomas. Es conocido por su capacidad para asistir en tareas que requieren comprensión y generación de lenguaje natural, incluyendo configuraciones técnicas y código.

Gemini Advanced 1.5 PRO de Google DeepMind es un modelo desarrollado por Google DeepMind que combina técnicas avanzadas de aprendizaje profundo y por refuerzo. Aunque sus especificaciones técnicas detalladas no son públicas, se sabe que está diseñado para abordar tareas complejas que requieren razonamiento y adaptación. Su entrenamiento en entornos interactivos mejora su capacidad para generar soluciones en contextos técnicos especializados.

LLaMA 3 de Meta AI es la tercera versión del modelo de lenguaje de Meta AI, está optimizado para ser eficiente sin sacrificar rendimiento. Disponible en distintos tamaños según la cantidad de parámetros, ha sido entrenado en un conjunto de datos multilingüe, lo que facilita su adaptación a tareas específicas. Su naturaleza de código abierto permite personalizaciones para necesidades particulares.

Claude 3.5 SONNET de Anthropic es un modelo enfocado en generar respuestas útiles y seguras, alineadas con principios éticos y de seguridad. Utiliza técnicas de aprendizaje por refuerzo con retroalimentación humana para minimizar sesgos y evitar contenido inapropiado. Aunque no se han divulgado detalles específicos sobre su arquitectura, se estima que cuenta con decenas de miles de millones de parámetros. Su enfoque en la seguridad lo hace relevante para aplicaciones en ciberseguridad.

Estos modelos fueron elegidos por su relevancia en la inteligencia artificial y su potencial para asistir en tareas técnicas especializadas. Representan diferentes enfoques y arquitecturas, lo que permite evaluar un espectro amplio de capacidades y limitaciones en el contexto de la ciberseguridad.

4.1.1.2. Definición de los Escenarios de Configuración

Se diseñaron cinco escenarios de configuración de políticas de seguridad, enfocándose en aspectos críticos de la administración de firewalls de nueva generación:

Políticas de Control de Acceso: Configuración de reglas para permitir o denegar tráfico basado en direcciones IP, puertos y protocolos específicos. Básicas en escenarios de riesgo de acceso no autorizado.

Políticas de Control de Aplicaciones: Establecimiento de reglas que gestionen el acceso a aplicaciones específicas, permitiendo un control granular del tráfico a nivel de aplicación. Utilizadas para el control de consumo y bloqueo de aplicaciones maliciosas o de alto riesgo.

Políticas Basadas en Identidad del Usuario: Implementación de políticas que restrinjan o permitan el acceso según la identidad del usuario, integrando servicios de autenticación y

directorios. Configuración aplicada para garantizar autenticación de acuerdo con los perfiles de los usuarios.

Políticas de Prevención de Intrusiones (IPS): Configuración de sistemas que detecten y prevengan actividades maliciosas o sospechosas en la red, aplicando firmas y técnicas de detección. Son fundamentales en la detección y respuesta de ataques basados en firmas o scripts.

4.1.1.3. Integración de Fabricantes Específicos

Para evaluar la capacidad de los LLMs de adaptarse a diferentes entornos tecnológicos, se incluyeron dos fabricantes líderes en soluciones de seguridad:

Fortinet: Reconocido por sus dispositivos FortiGate, que ofrecen funcionalidades avanzadas de firewall y seguridad unificada. Es uno de los fabricantes con mayor presencia en las industria por la accesibilidad de sus soluciones para grandes y pequeñas empresas.

Se tomo como referente la versión recomendada por el fabricante FortiOS 7.4.6, esta cuenta con la documentación disponible de manera pública y estará vigente hasta el ultimo trimestre del 2026.

Check Point: Destacado por sus soluciones de seguridad integrales y su enfoque en la prevención de amenazas. Este fabricante se ha mantenido en el top del cuadrante de lideres de Gartner, organización que caracteriza la eficacia de soluciones de seguridad.

Para este estudio se seleccionó su versión R81.20 del sistema operativo GAIA, la cual es la recomendada por el fabricantes y que permite la integración de una API para la creación directa de configuración por código, demás que cuenta con una extensa documentación de configuración.

Cada escenario se presentó a los LLMs en tres modalidades:

Escenario Genérico: Se describió el escenario sin mencionar ningún fabricante, solicitando una configuración aplicable a cualquier firewall de nueva generación.

Escenario Específico para Fortinet: Se indicó que la configuración debía aplicarse a dispositivos Fortinet, proporcionando cuando fue posible la guía de administración oficial, "FortiOS-7.4.6-Administration_Guide.pdf".

Escenario Específico para Check Point: Se solicitó una configuración adaptada a dispositivos Check Point, también suministrando la documentación oficial, "CP_R81.20_Gaia_AdminGuide.pdf", si el modelo lo permitía.

Esta estrategia permitió evaluar si los LLMs pueden generar configuraciones específicas para distintos fabricantes y cómo influye el suministro de documentación adicional en la calidad de las respuestas.

4.1.1.4. Preparación de las Interacciones con los LLMs

Se diseñaron instrucciones(promt) específicos para cada escenario y modalidad, asegurando claridad y consistencia en las solicitudes.

Se crearon promt genéricos que describían el escenario y solicitaban una configuración detallada sin referirse a ningún fabricante, posteriormente a la prueba base, promt específicos que incluían la indicación del fabricante y, cuando era posible, se adjuntaba la documentación oficial relevante.

Control de Acceso:

Escenario: Se necesita configurar una política de control de acceso en un firewall de nueva generación. La política debe permitir que los administradores de servidores, ubicados en la red interna 192.168.100.0/24, accedan al servidor web de la compañía con dirección IP 10.10.10.20. Los servicios que deben estar permitidos son Telnet, SSH, HTTP y HTTPS.

Promt: "Se requiere configurar una política en un firewall de nueva generación que permita a los administradores de servidores de la red 192.168.100.0/24(ADMIN_NET) acceder al servidor web con IP 10.10.10.20(SRV_ONE). Los servicios que deben estar permitidos son: Telnet (puerto 23/TCP), SSH (puerto 22/TCP), HTTP (puerto 80/TCP) y HTTPS (puerto 443/TCP).

Proporciona la instrucción(código) para crear esta política, asume que los objetos necesarios ya están creados en la base de datos del firewall"

Especificando fabricante: "Proporciona la instrucción(código) para crear esta política en un dispositivo [Fortinet / Check Point], siguiendo el procedimiento que se encuentra en el documento adjunto (guía de administración del fabricante) y asumiendo que los objetos necesarios ya están creados en la base de datos del firewall."

Control de Aplicaciones

Escenario: Se necesita configurar una política de control de aplicaciones en un firewall de nueva generación. La política debe gestionar el acceso a aplicaciones específicas para permitir un control granular del tráfico a nivel de aplicación. Específicamente, se requiere:

Bloquear el acceso a aplicaciones de intercambio de archivos P2P, como BitTorrent y eMule, debido a su alto riesgo de seguridad.

Permitir el acceso a aplicaciones de colaboración empresarial, como Microsoft Teams y Slack, necesarias para las operaciones diarias.

La política debe aplicarse a todos los usuarios en la red interna de la compañía.

Promt: "Se requiere configurar una política en un firewall de nueva generación que realice lo siguiente:

Bloquear el acceso a aplicaciones de intercambio de archivos P2P, incluyendo Torrent.

Permitir el acceso a aplicaciones de colaboración empresarial, como Microsoft Teams y Slack.

La política debe aplicarse a todos los usuarios en la red interna.

Proporciona la instrucción(código) para crear esta política, asume que los objetos necesarios ya están creados en la base de datos del firewall "

Especificando fabricante:

" Especificando fabricante: "Proporciona la instrucción(código) para crear esta política en un dispositivo [Fortinet / Check Point], siguiendo el procedimiento que se encuentra en el documento adjunto (guía de administración del fabricante) y asumiendo que los objetos necesarios ya están creados en la base de datos del firewall."

Basadas en Identidad de Usuario

Escenario: Se necesita configurar una política en un firewall de nueva generación que restrinja y permita el acceso a recursos según la identidad de los usuarios, integrando servicios de autenticación y directorios. Específicamente, se requiere lo siguiente:

Permitir que los usuarios del departamento de Finanzas, identificados en el Directorio Activo (Active Directory) del dominio empresa.local, accedan al servidor de base de datos financiero con IP 10.20.30.40 a través del puerto 1521/TCP (Oracle Database).

Denegar el acceso a este servidor para todos los demás usuarios.

La autenticación debe integrarse con el Active Directory existente para garantizar que solo los usuarios autorizados puedan acceder al recurso según su perfil.

Promt: "Se requiere configurar una política en un firewall de nueva generación que cumpla con las siguientes especificaciones:

Permitir que los usuarios del departamento de Finanzas, identificados en el Active Directory del dominio empresa.local, accedan al servidor de base de datos con IP 10.20.30.40 a través del puerto 1521/TCP (Oracle Database).

Denegar el acceso a este servidor para todos los demás usuarios.

La política debe integrar servicios de autenticación con el Active Directory para validar la identidad de los usuarios.

Proporciona la instrucción(código) para crear esta política, asume que los objetos necesarios ya están creados en la base de datos del firewall."

Especificando fabricante: “Proporciona la instrucción(código) para crear esta política en un dispositivo [Fortinet / Check Point], siguiendo el procedimiento que se encuentra en el documento adjunto (guía de administración del fabricante) y asumiendo que los objetos necesarios ya están creados en la base de datos del firewall.”

Prevención de Intrusiones (IPS)

Escenario: Se necesita configurar una política de prevención de intrusiones (IPS) en un firewall de nueva generación. La política debe detectar y prevenir actividades maliciosas o sospechosas en la red, aplicando firmas y técnicas de detección. Específicamente, se requiere:

Habilitar la detección y prevención de ataques conocidos como SQL Injection y Cross-Site Scripting (XSS) en el tráfico web dirigido a los servidores internos con IP 10.10.10.20 y 10.10.10.21.

Aplicar las firmas de IPS correspondientes a estos tipos de ataques.

Configurar el sistema para que genere alertas y bloquee automáticamente cualquier tráfico malicioso detectado.

Promt: "Se requiere configurar una política de prevención de intrusiones (IPS) en un firewall de nueva generación que cumpla con las siguientes especificaciones:

Habilitar la detección y prevención de ataques de SQL Injection y Cross-Site Scripting (XSS) en el tráfico web dirigido a los servidores con IP 10.10.10.20 y 10.10.10.21.

Aplicar las firmas de IPS correspondientes para estos ataques.

Configurar el sistema para que genere alertas y bloquee automáticamente el tráfico malicioso detectado.

Proporciona la instrucción(código) para crear esta política, asume que los objetos necesarios ya están creados en la base de datos del firewall."

Especificando fabricante: “Proporciona la instrucción(código) para crear esta política en un dispositivo [Fortinet / Check Point], siguiendo el procedimiento que se encuentra en el documento adjunto (guía de administración del fabricante) y asumiendo que los objetos necesarios ya están creados en la base de datos del firewall.”

4.1.1.5. Definición de Criterios de Evaluación.

Con el fin de verificar en qué medida los resultados generados por los modelos de lenguaje se alinean con los estándares y buenas prácticas en materia de seguridad de la información, así como con las directrices de los fabricantes, se definieron criterios específicos para cada

escenario con los cuales se evaluaría su nivel de cumplimiento. Estos criterios abarcan configuraciones básicas que deben incorporarse en las políticas para mantener un control seguro de la información.

Para la definición de los criterios de evaluación, se tomaron como base tres documentos fundamentales que proporcionan estándares y buenas prácticas ampliamente reconocidas en la industria. En primer lugar, el CIS Check Point Firewall Benchmark v1.1.0 (Center for Internet Security, 2020), que ofrece una guía detallada para establecer configuraciones seguras en dispositivos de firewall Check Point. En segundo lugar, el CIS Fortigate 7.0.x Benchmark v1.3.0 (Center for Internet Security, 2024), el cual establece lineamientos específicos para fortalecer la seguridad de dispositivos FortiGate. Finalmente, se utilizó el NIST Special Publication 800-41 Revision 1 (Scarfone & Hoffman, 2009), que proporciona recomendaciones exhaustivas sobre políticas y tecnologías de firewall para proteger las redes organizacionales. Estos documentos permitieron identificar configuraciones clave que garantizaran la alineación con las mejores prácticas en seguridad de la información y la mitigación de riesgos asociados.

Criterios de Evaluación para Políticas de Control de Acceso

A continuación se describen los criterios establecidos para las políticas de control de acceso, incluyendo para cada uno su descripción, justificación y un ejemplo de implementación tanto en dispositivos Check Point como en Fortinet.

- **Definición de Zonas o Interfaces Explícitas**

  Descripción: Este criterio evalúa si las configuraciones de control de acceso especifican de manera clara las zonas de origen y destino en Check Point, o las interfaces de entrada y salida en dispositivos Fortinet. El objetivo es segmentar eficazmente el tráfico entre diferentes áreas de la red, controlando su flujo y previniendo el acceso no autorizado.

  Justificación: La definición explícita de zonas o interfaces resulta fundamental para garantizar una segmentación adecuada. Al delimitar áreas críticas y no críticas, se reducen las posibilidades de que el tráfico no autorizado transite por la red,

disminuyendo así la superficie de ataque. Además, esta práctica asegura que las políticas reflejen el diseño de seguridad establecido para la infraestructura de red.

Ejemplo: En Check Point: Configurar una regla con layer: Network, que indique source: INTERNAL_ZONE y destination: EXTERNAL_ZONE.

En Fortinet: Definir el tráfico de origen y destino a través de srcintf: port1 (red interna) y dstintf: port2 (red externa).

- **Direcciones Explícitas**

  Descripción: Este criterio valora el uso de direcciones específicas para las fuentes y destinos de la política de control de acceso, en lugar de recurrir a valores genéricos como “ANY”. Se busca así asegurar una mayor precisión y minimizar la posibilidad de que se permita tráfico no deseado.

  Justificación: La especificación de direcciones concretas reduce significativamente los riesgos de accesos no controlados y mejora la trazabilidad de los eventos de seguridad. Al utilizar definiciones genéricas (por ejemplo, “ANY”), se incrementa la probabilidad de exponer la red a amenazas potenciales, dificultando además el proceso de auditoría y monitoreo.

  Ejemplo: En Check Point: Definir la regla con source: ADMIN_NET (subred administrativa) y destination: SRV_ONE (servidor crítico).

  En Fortinet: Configurar la política con set srcaddr ADMIN_NET y set dstaddr SRV_ONE, estableciendo direcciones explícitas de origen y destino.

- **Acción Definida**

  Descripción: Se verifica que cada regla especifique de forma inequívoca la acción que debe tomarse ante el tráfico evaluado, ya sea permitir (Accept), denegar (Deny) o descartar (Drop). Así, se evita la ambigüedad y se garantiza un comportamiento coherente del firewall.

  Justificación: Definir claramente la acción para cada flujo de tráfico es esencial para alinear las políticas de seguridad con los objetivos de la organización. Una configuración que no indique la acción deseada puede dar lugar a inconsistencias y vulnerabilidades, afectando la eficacia general de la protección de la red.

Ejemplo: En Check Point: Una regla que indica Action: Accept para permitir tráfico entre zonas específicas.

En Fortinet: Configuración de la regla con set action accept, estableciendo de forma explícita el permiso para el tráfico correspondiente.

- **Habilitación de Registros (Logging)**

  Descripción: Este criterio se centra en la activación de registros para monitorear y analizar los eventos de seguridad. Se recomienda generar logs tanto de los accesos permitidos como de los denegados, especialmente en las reglas críticas.

  Justificación: La habilitación de registros es indispensable para detectar y comprender incidentes de seguridad. Estos registros proporcionan evidencia que puede emplearse en auditorías e investigaciones forenses, además de posibilitar la identificación de comportamientos anómalos o intentos de acceso no autorizado. El monitoreo continuo permite mejorar las políticas de seguridad y reforzar la protección de la red.

  Ejemplo: En Check Point: Activar la opción Enable Logging en todas las reglas relevantes, a fin de registrar todos los eventos críticos.

  En Fortinet: Configurar set logtraffic all para generar registros de todo el tráfico (permitido o denegado).

Criterios de Evaluación para las Políticas de Control de Aplicaciones

A continuación, se detallan los criterios establecidos para evaluar las políticas de control de aplicaciones, incluyendo su descripción, justificación y ejemplos de implementación tanto en Check Point como en Fortinet.

- **Identificación de Aplicaciones Específicas y Categorías**

  Descripción: Este criterio examina si las políticas contemplan la identificación de aplicaciones particulares (por ejemplo, YouTube, WhatsApp) o categorías de aplicaciones (como redes sociales, streaming, mensajería). De esta forma, se ejerce un control más preciso sobre el tráfico generado por cada categoría o aplicación.

Justificación: La identificación detallada de aplicaciones y categorías resulta fundamental para priorizar aquellas que son críticas para el negocio y, al mismo tiempo, restringir o bloquear las que puedan suponer un riesgo de seguridad o afectar la productividad. Un ejemplo común es bloquear el acceso a redes sociales durante el horario laboral para incrementar la eficiencia de los colaboradores.

Ejemplo: En Check Point: La política reconoce Category: Social Media e incluye aplicaciones específicas como Facebook e Instagram.

En Fortinet: Se configura un filtro de aplicaciones que contempla Application Category: Streaming Media para ejercer control sobre contenido de streaming.

- **Definición de Acciones Específicas**

  Descripción: Las políticas deben establecer de manera explícita las acciones (permitir, bloquear o restringir) aplicables a cada aplicación o categoría, contemplando posibles restricciones de horario o límites de ancho de banda.

  Justificación: La determinación de acciones concretas para cada aplicación contribuye a gestionar el tráfico de forma segura y eficiente, ajustándose a las necesidades particulares de la organización. Por ejemplo, limitar el uso de YouTube fuera del horario laboral puede evitar la saturación de la red durante las horas de mayor actividad.

  Ejemplo: En Check Point: Una regla específica señala Action: Block para la categoría Streaming Media durante el horario laboral.

  En Fortinet: Se implementa la configuración set application-block app-group "SocialMediaGroup" en conjunto con set schedule "work_hours" para restringir el uso de aplicaciones de redes sociales en horas de trabajo.

- **Monitoreo y Registro del Tráfico de Aplicaciones**

  Descripción: Este criterio determina si las políticas contemplan el monitoreo continuo y el registro de eventos de tráfico relacionados con las aplicaciones. De este modo, se facilita la auditoría y la detección de anomalías.

  Justificación: La supervisión constante permite identificar comportamientos inusuales que podrían señalar incidentes de seguridad. Además, el registro de actividades resulta

clave en auditorías de cumplimiento y análisis de comportamiento, ya que brinda datos concretos sobre el uso real de aplicaciones en la red.

Ejemplo: En Check Point: Se habilita la opción Enable Logging para registrar el uso de aplicaciones específicas, como WhatsApp.

En Fortinet: Se configura set logtraffic all con el fin de registrar todo el tráfico asociado a un Application Group determinado.

- **Soporte para Inspección de Tráfico Encriptado (SSL/SSH)**

  Descripción: Este criterio evalúa si la configuración de las políticas incluye la inspección de tráfico encriptado, como SSL o SSH, lo cual permite analizar aplicaciones que emplean cifrado avanzado.

  Justificación: Dado que muchas aplicaciones modernas utilizan cifrado para sus comunicaciones, la capacidad de inspeccionar este tráfico es esencial para identificar y gestionar adecuadamente su uso sin reducir la seguridad de la red. Así, se previenen posibles vulnerabilidades derivadas de una falta de visibilidad en el tráfico cifrado.

  Ejemplo: En Check Point: Se habilita la inspección de SSL para descifrar y analizar el tráfico HTTPS.

  En Fortinet: La configuración se realiza mediante set ssl-ssh-profile "deep-inspection", permitiendo la inspección profunda del tráfico encriptado.

- **Configuración Granular de Políticas**

  Descripción: Este criterio observa si las políticas permiten niveles de configuración detallados, incluyendo límites específicos por aplicación, categoría, horario, ancho de banda o usuario.

  Justificación: La posibilidad de establecer configuraciones granulares brinda un control más eficiente y ajustado a las necesidades de cada entorno. Al definir reglas específicas para diferentes grupos de usuarios o aplicaciones, se minimizan los riesgos asociados a políticas demasiado amplias que podrían ser objeto de explotación maliciosa.

  Ejemplo: En Check Point: Una regla establece restricciones de ancho de banda para YouTube y define horarios específicos de acceso.

  En Fortinet: Se aplica set app-control "high" para reforzar el nivel de control, además de configurar límites de ancho de banda mediante set traffic-shaper.

Criterios de Evaluación para Políticas Basadas en Control de Identidad

A continuación, se presentan los criterios definidos para evaluar políticas basadas en control de identidad, ofreciendo para cada uno su descripción, justificación y ejemplos de aplicación en diferentes plataformas.

- **Integración con Servicios de Autenticación y Directorios**

  Descripción: Las políticas deben conectarse con sistemas de autenticación como Active Directory (AD) u otros servicios equivalentes (LDAP, RADIUS), a fin de centralizar la gestión de usuarios y perfiles. Esto permite que los cambios efectuados en los directorios se reflejen automáticamente en las políticas de seguridad.

  Justificación: La integración con servicios de autenticación garantiza que las políticas se apliquen de manera uniforme y se adapten a cualquier modificación en los perfiles de los usuarios (por ejemplo, cambios de roles o revocación de permisos). De esta forma, se simplifica la administración de identidades y se fortalece el control de acceso, ya que solo los usuarios autorizados podrán acceder a los recursos.

  Ejemplo: En Check Point: Configurar la integración con AD empleando Authentication Settings e Identity Awareness.

  En Fortinet: Utilizar set ldap-server para enlazar con el servidor LDAP corporativo.

- **Restricción de Acceso Según Grupos de Usuarios**

  Descripción: Las políticas deben contemplar la posibilidad de restringir el acceso a recursos en función de los grupos de usuarios establecidos en el sistema de autenticación (por ejemplo, “Finanzas”, “Recursos Humanos”).

  Justificación: La segmentación del acceso según grupos de usuarios promueve el principio del menor privilegio, otorgando únicamente los permisos indispensables para las funciones de cada rol. Con ello, se reducen riesgos de accesos innecesarios y se mantiene un mejor control sobre los recursos críticos.

  Ejemplo: En Check Point: Establecer una regla que permita el acceso al servidor Finance_DB únicamente al grupo Finance_Users.

  En Fortinet: Configurar las políticas con set user-group "HR_Group" para restringir el acceso a determinados recursos.

- **Permisos Explícitos y Acciones Claras**

  Descripción: Las políticas deben definir de manera explícita los permisos de acceso asignados a cada usuario o grupo, indicando claramente las acciones permitidas o denegadas.

  Justificación: Al especificar con claridad las acciones autorizadas o bloqueadas, se evitan interpretaciones ambiguas que podrían derivar en accesos no deseados. Además, esto brinda a los usuarios un entendimiento preciso de los recursos a los que pueden acceder y en qué condiciones.

  Ejemplo: En Check Point: Una política define Action: Allow para el grupo IT_Admins en el servidor de administración, mientras que establece Action: Deny para otros usuarios.

  En Fortinet: Utilizar set action deny para restringir el acceso a quienes no pertenecen al grupo autorizado.

- **Seguridad en la Autenticación**

  Descripción: Las políticas deben respaldar el uso de protocolos seguros, como LDAP sobre TLS/SSL, para proteger las credenciales de los usuarios durante los procesos de autenticación y prevenir interceptaciones.

  Justificación: El cifrado de las credenciales en tránsito impide su interceptación y reduce la posibilidad de ataques de suplantación de identidad o robo de contraseñas. Esta práctica fortalece la confianza en la infraestructura de autenticación y en la eficacia de las políticas de control de identidad.

  Ejemplo: En Check Point: Configurar la autenticación segura mediante la opción Use TLS for LDAP Connections.

  En Fortinet: Activar la encriptación de la comunicación con set secure enable en las configuraciones de LDAP.

Criterios de Evaluación para Políticas de IPS (Intrusion Prevention System)

A continuación, se presentan los principales criterios de evaluación para la implementación de políticas de IPS, junto con su descripción, justificación y ejemplos prácticos en entornos Check Point y Fortinet.

- **Habilitación de Firmas Específicas**

  Descripción: La política de IPS debe incluir firmas específicas para detectar ataques comunes, tales como SQL Injection y Cross-Site Scripting (XSS). De este modo, se logra identificar y mitigar patrones de amenazas ampliamente conocidos en el entorno de aplicaciones web.

  Justificación: Los ataques SQL Injection y XSS se cuentan entre los vectores más frecuentes para comprometer aplicaciones y bases de datos. Al habilitar firmas específicas para estas amenazas, se protege de forma proactiva la infraestructura interna y se bloquean intentos de explotar vulnerabilidades reconocidas.

  Ejemplo: En Check Point: Configuración de una firma IPS específica para SQL_Injection y XSS.

  En Fortinet: Habilitación de la firma mediante set signature SQL_INJECTION enable.

- **Configuración de Acciones (Bloqueo y Alerta)**

  Descripción: La política debe especificar acciones concretas para el tráfico malicioso, tales como bloquearlo de inmediato y generar alertas en tiempo real. Esto asegura una respuesta rápida ante incidentes y facilita el monitoreo continuo de la red.

  Justificación: El bloqueo de tráfico malicioso elimina la posibilidad de que un ataque en curso afecte la integridad de la red. Por otra parte, las alertas suministran información valiosa a los administradores, permitiéndoles supervisar la actividad anómala y tomar medidas adicionales cuando sea necesario.

  Ejemplo: En Check Point: Uso de Action: Prevent para bloquear y Log: Alert para generar registros del evento.

  En Fortinet: Configuración de set action block y set log enable, bloqueando y registrando los intentos de intrusión.

- **Aplicación en Objetivos Específicos**

  Descripción: La política de IPS debe implementarse únicamente en los servidores internos considerados críticos, como aquellos con direcciones IP específicas (por ejemplo, 10.10.10.20 y 10.10.10.21), para maximizar la efectividad y minimizar el impacto en el rendimiento.

Justificación: Enfocar la protección en los sistemas más valiosos y críticos reduce el consumo innecesario de recursos en áreas de la red con menor relevancia, optimizando el desempeño del IPS. Esto garantiza que se prioricen los activos esenciales de la organización ante posibles ataques.

Ejemplo: En Check Point: Definición de la política con Source: Any, Destination: 10.10.10.20, 10.10.10.21.

En Fortinet: Configuración que utiliza set dstaddr "Server_Group", donde Server_Group incluye las IP críticas.

Uso de Técnicas Avanzadas de Detección

Descripción: La política debe integrar métodos de detección basados tanto en firmas como en análisis de comportamiento anómalo. Este enfoque dual permite identificar amenazas conocidas y nuevas variantes que aún no cuentan con una firma específica.

Justificación: Al combinar técnicas de detección por firmas y detección de anomalías, se fortalece el nivel de protección brindado por el IPS. Mientras que las firmas permiten reaccionar a ataques documentados, la detección de comportamientos atípicos aporta un enfoque preventivo ante amenazas emergentes.

Ejemplo: En Check Point: Activación de análisis combinado a través de un Threat Prevention Profile que incluye firmas y monitoreo de comportamiento.

En Fortinet: Configuración avanzada de detección con set ips-sensor "Advanced_Sensor", que integra firmas y detección de anomalías.

Calificación de Cumplimiento

Alto: Cumple con todos los criterios establecidos.

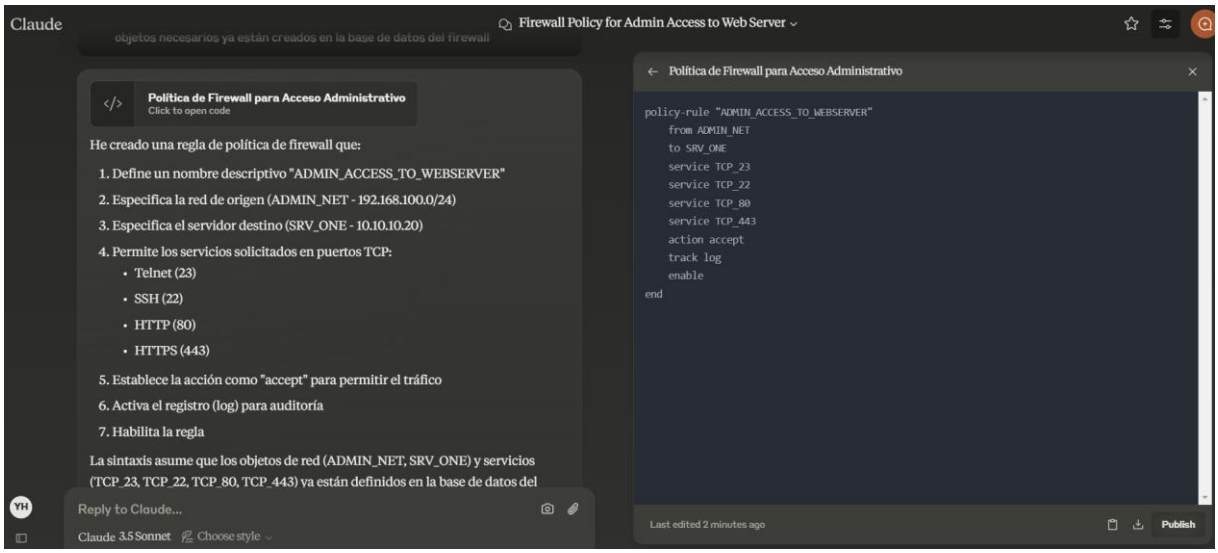
Medio: Cumple con más de la mitad de los criterios.

Bajo: Cumple con menos de la mitad de los criterios evaluados.

4.1.1.6. Obtención y Evaluación de las Configuraciones Generadas LLMs

El presente apartado describe el procedimiento llevado a cabo para la generación de configuraciones de firewall mediante los Modelos de Lenguaje (LLMs) seleccionados, así como la evaluación de su calidad y conformidad con los estándares y buenas prácticas definidas. Este proceso se enmarca en el piloto experimental descrito en secciones anteriores, cuyo objetivo principal consiste en verificar la eficacia de los LLMs en escenarios reales de configuración de políticas de seguridad.

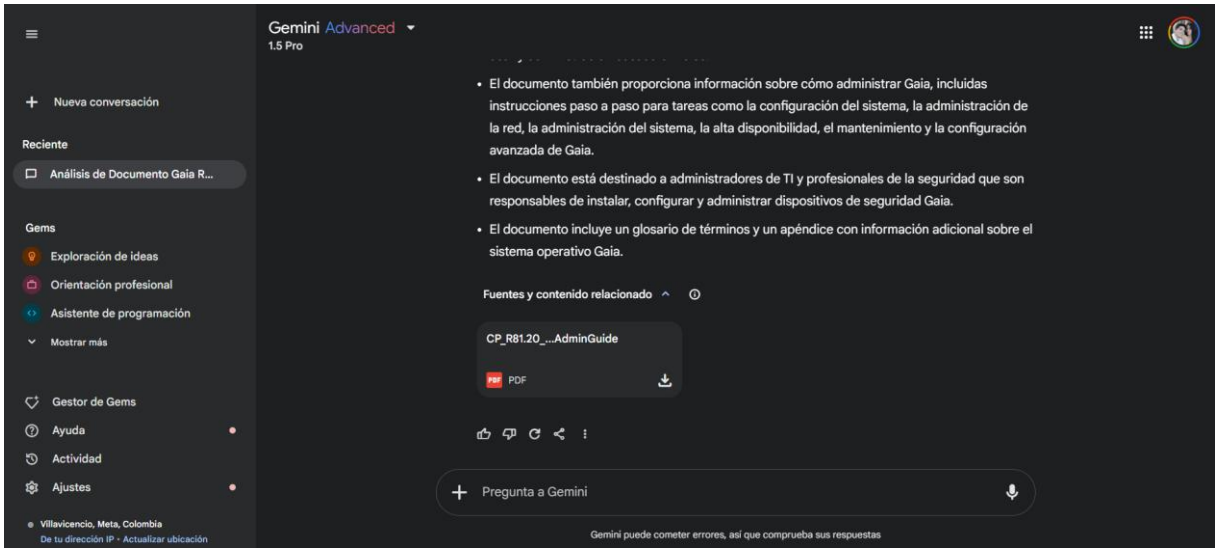
Ilustración 1 Interacción del escenario de política de acceso con el LLM Claude.



Fuente. Elaboración propia.

En primer lugar, se procedió a interactuar con cada LLM de forma independiente, siguiendo la metodología expuesta en la Sección 3.3. Para cada uno de los cuatro escenarios planteados (Control de Acceso, Control de Aplicaciones, Políticas Basadas en Identidad de Usuario y Prevención de Intrusiones –IPS–), se elaboraron prompts específicos que describían la tarea a realizar. Dichos prompts se presentaron en tres modalidades: una genérica (sin fabricante) y dos particulares (Fortinet y Check Point). Cuando la plataforma lo permitió, se incluyó además la documentación oficial asociada al fabricante, con el fin de reforzar la exactitud de las respuestas y la adaptación a cada entorno tecnológico.

Ilustración 2 Interacción con el LLM Gemini cargando el Admin Guide CheckPoint.



Fuente. Elaboración propia.

La ejecución secuencial de todos los escenarios por cada LLM tuvo como propósito evitar interferencias entre modelos y asegurar la obtención de resultados independientes. Dado que algunos LLMs permiten conservar un historial de interacciones, se habilitaron chats temporales o chats nuevos en cada prueba, de modo que no se arrastraran datos ni instrucciones de experimentos previos. Asimismo, las guías de administración de cada fabricante se dividieron en secciones manejables cuando superaban un determinado número de páginas, garantizando que los LLMs pudieran procesar la información de manera coherente.

Ilustración 3 Interacción el LLM CHATGPT con Chat Temporal habilitado.

Fuente. Elaboración propia.

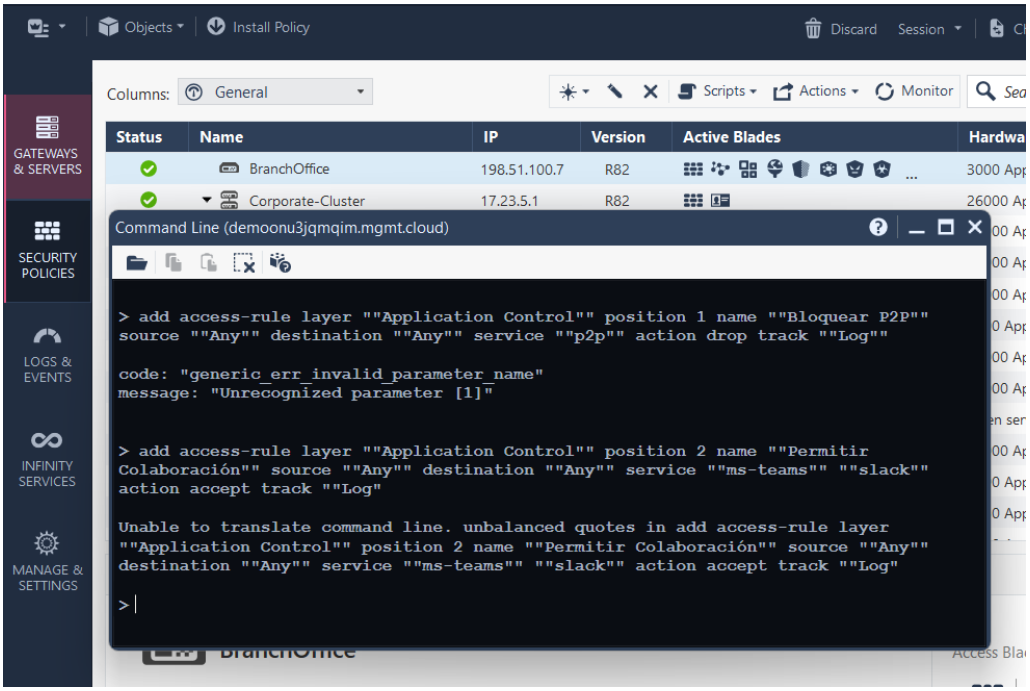
Una vez obtenidas las configuraciones sugeridas por cada modelo, se llevó a cabo un registro ordenado de la información en una matriz de resultados. En esta matriz se incluyeron tanto los bloques de código o comandos de firewall generados como los comentarios y explicaciones adicionales proporcionados por la IA. De este modo, fue posible contar con un insumo centralizado y estandarizado para la evaluación posterior.

Ilustración 4 Interacción del escenario de política de acceso con el LLM Meta AI.

Fuente. Elaboración propia.

La siguiente etapa consistió en una validación funcional preliminar, orientada únicamente a los escenarios en los que se especificó explícitamente el fabricante. Para Check Point se empleó la demostración de SmartConsole, con acceso a la API y la posibilidad de ejecutar directamente los comandos recomendados; en el caso de Fortinet, se configuró una máquina virtual en la que se ingresaron las secuencias de comandos proporcionadas por la IA. Esta validación permitió comprobar la aplicabilidad y operatividad de las configuraciones, identificando posibles errores de sintaxis u omisiones críticas en los parámetros requeridos para cada política de seguridad.

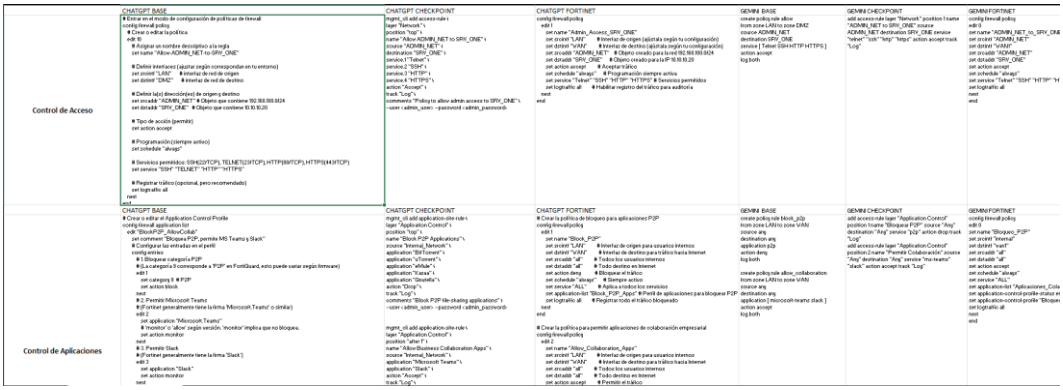
Ilustración 5 SmartConsole Checkpoint API dentro de un ambiente DEMO



Fuente. Elaboración propia.

Con el fin de apreciar el nivel de alineación de las configuraciones con los estándares establecidos, se contrastaron los resultados obtenidos frente a los CIS Benchmarks específicos de Fortinet y Check Point, así como las directrices generales del NIST SP 800-41. Para ello, se definieron criterios de evaluación basados en los lineamientos de seguridad sugeridos por estos marcos de referencia. Dichos criterios se agruparon en hojas de cálculo, clasificando los hallazgos de cada escenario y fabricante, estos criterios se encuentran detallados en el anexo matriz de resultados. De este modo, se determinaron el grado de conformidad, la existencia de errores u omisiones relevantes y la coherencia general de cada política propuesta.

Ilustración 6 Sección del anexo Matriz de Resultados



Fuente. Elaboración propia.

Finalmente, se realizó una correlación entre la validación funcional y la evaluación de cumplimiento de estándares, asignando a cada modelo una calificación o valoración que refleja tanto la operatividad de las configuraciones como su apego a las buenas prácticas de la industria. Este proceso de análisis permitió identificar patrones de comportamiento comunes en los LLMs, así como las áreas donde se advirtieron mayores divergencias.

4.1.2. Descripción de los resultados

En esta sección se exponen de manera objetiva los resultados obtenidos tras la validación funcional y la evaluación de conformidad de las configuraciones generadas por los LLMs (ChatGPT, Gemini, Meta IA y Claude) en los diferentes escenarios de firewall. Se presentan los datos cuantitativos organizados en tablas de resumen.

La Tabla 1 resume el funcionamiento de las configuraciones propuestas por cada IA (ChatGPT, Gemini, Meta IA y Claude) en cuatro categorías seleccionadas:

Tabla 1 Evaluación de funcionamiento.

AI	FABRICANTE	ACCESO	APLICACION	IDENTIDAD	IPS	%Exito
CHATGPT	Checkpoint	1	1	1	1	75%
	Fortinet	1	1	0	0	
GEMINI	Checkpoint	1	0	0	0	38%
	Fortinet	1	0	0	1	
META IA	Checkpoint	1	0	0	0	38%
	Fortinet	1	1	0	0	
CLAUDE	Checkpoint	1	0	0	0	50%
	Fortinet	1	0	1	1	
		100%	38%	25%	38%	

Fuente: Elaboración propia con base en los resultados del piloto.

Los valores presentados (1 = Sí, 0 = No) indican si la configuración proporcionada por la IA resultó funcional en el entorno de cada fabricante (Check Point y Fortinet). En la última columna de la tabla se incluye el porcentaje de éxito global por IA y en la última fila por escenario.

La Tabla 2 muestra el grado de cumplimiento en Control de Acceso, considerando criterios como la definición explícita de zonas, la especificación de direcciones, la definición de acciones y la habilitación de registros. Cada criterio se evalúa con 1 (cumple) o 0 (no cumple), y se presenta un subtotal para cada plataforma junto con un total general por IA.

Tabla 2 Cumplimiento Control de Acceso

CUMPLIMIENTO CONTROL DE ACCESO								
AI	PLATAFORMA	NIVEL DE CUM	Explicit Zones	Explicit Addre	Action Define	Logging Enable	Subtotal	Total
CHATGPT	Checkpoint	Alto	1	1	1	1	4	
	Fortinet	Alto	1	1	1	1	4	
	BASE	Alto	1	1	1	1	4	12
GEMINI	Checkpoint	Alto	1	1	1	1	4	
	Fortinet	Alto	1	1	1	1	4	
	BASE	Alto	1	1	1	1	4	12
META IA	Checkpoint	Medio	0	1	1	1	3	
	Fortinet	Alto	1	1	1	1	4	
	BASE	Medio	0	1	1	1	3	10
CLAUDE	Checkpoint	Alto	1	1	1	1	4	
	Fortinet	Alto	1	1	1	1	4	
	BASE	Medio	0	1	1	1	3	11
	Total		9	12	12	12		

Fuente: Elaboración propia con base en los resultados del piloto.

La Tabla 3 muestra el grado de cumplimiento de los lineamientos de Control de Aplicaciones para cada IA y plataforma. Se incluyen varios criterios, como la Identificación de Aplicaciones, Acciones Específicas, Monitoreo y Registro, Soporte para inspección de tráfico encriptado y Configuración Granular.

Tabla 3 Cumplimiento Control de Aplicaciones

CUMPLIMIENTO CONTROL DE APP									
AI	PLATAFORMA	NIVEL DE CUMPLIMIENTO	Identificacion de Aplicaciones	Acciones Especificas	Monitoreo y Registro	Soporte para inspeccion trafico encriptado	Configuracion Granular	Subtotal	Total
CHATGPT	Checkpoint	Medio	1	1	1	0	1	4	
	Fortinet	Medio	1	1	1	0	0	3	
	BASE	Medio	1	1	1	0	1	4	11
GEMINI	Checkpoint	Medio	1	1	1	0	0	3	
	Fortinet	Alto	1	1	1	1	1	5	
	BASE	Medio	1	1	1	0	0	3	11
METAIA	Checkpoint	Medio	1	1	1	0	0	3	
	Fortinet	Medio	1	1	1	0	0	3	
	BASE	Medio	1	1	1	0	0	3	9
CLAUDE	Checkpoint	Medio	1	1	1	0	1	4	
	Fortinet	Alto	1	1	1	1	1	5	
	BASE	Medio	1	1	1	0	0	3	12
		Total	12	12	12	2	5		

Fuente: Elaboración propia con base en los resultados del piloto.

En la columna “Subtotal” se encuentran los valores consolidados de cada IA según la plataforma, y en “Total” se observa el puntaje sumatorio global.

La Tabla 4 recoge el cumplimiento en los aspectos de Políticas basadas en Identidad, en los cuales se verifican criterios como Integración con Active Directory u otros directorios, uso de Grupos y Perfiles Explícitos, y la configuración de Autenticación. Cada criterio está marcado con 1 (cumple) o 0 (no cumple), agrupándose el subtotal y total para cada IA y plataforma.

Tabla 4 Cumplimiento Políticas Basadas en Identidad

CUMPLIMIENTO IDENTIDAD								
AI	PLATAFORMA	NIVEL DE CUMPLIMIENTO	Integracion	Grupos	Explicitos	Autenticacion	Subtotal	Total
CHATGPT	Checkpoint	Medio	0	1	1	0	2	
	Fortinet	Medio	1	1	1	0	3	
	BASE	Medio	1	1	1	0	3	8
GEMINI	Checkpoint	Medio	0	1	1	0	2	
	Fortinet	Medio	1	1	1	0	3	
	BASE	Medio	1	1	1	0	3	8
META IA	Checkpoint	Medio	1	1	1	0	3	
	Fortinet	Medio	1	1	1	0	3	
	BASE	Medio	1	1	1	0	3	9
CLAUDE	Checkpoint	Medio	1	1	1	0	3	
	Fortinet	Medio	1	1	1	0	3	
	BASE	Medio	1	1	1	0	3	9
		Total	10	12	12	0		

Fuente: Elaboración propia con base en los resultados del piloto.

La Tabla 5 muestra los resultados del Cumplimiento en IPS, evaluando elementos como la inclusión de Firmas, la capacidad de Bloqueo, la definición de Objetivos de protección y el nivel de Protección Avanzada ofrecido por la configuración generada. Se presentan los subtotales y el total para cada IA.

Tabla 5 Cumplimiento Políticas IPS

CUMPLIMIENTO IPS								
AI	PLATAFORMA	NIVEL DE CUMPLIMIENTO	Firmas	Bloqueo	Objetivos	Avanzada	Subtotal	Total
CHATGPT	Checkpoint	Medio	1	1	1	0	3	
	Fortinet	Medio	1	1	1	0	3	
	BASE	Alto	1	1	1	1	4	10
GEMINI	Checkpoint	Medio	1	1	0	0	2	
	Fortinet	Medio	1	1	1	0	3	
	BASE	Medio	1	1	1	0	3	8
META IA	Checkpoint	Medio	1	1	1	0	3	
	Fortinet	Medio	1	1	1	0	3	
	BASE	Medio	1	1	0	0	2	8
CLAUDE	Checkpoint	Medio	1	1	1	0	3	
	Fortinet	Medio	1	1	1	0	3	
	BASE	Medio	1	1	1	0	3	9
		Total	12	12	10	1		

Fuente: Elaboración propia con base en los resultados del piloto.

La Tabla 6 condensa los porcentajes globales de cumplimiento por IA y por escenario (Acceso, Aplicación, Identidad e IPS). Se muestra el valor total obtenido en cada categoría y el porcentaje respectivo, facilitando la comparación en una vista sintetizada.

Tabla 6 Consolidad Porcentaje de Cumplimiento

PORCENTAJE DE CUMPLIMIENTO					
AI	CUMPLIMIENTO PROMEDIO	ACCESO	APLICACION	IDENTIDAD	IPS
CHATGPT	41	12	11	8	10
	80.4%	100.0%	73.3%	66.7%	83.3%
GEMINI	39	12	11	8	8
	76.5%	100.0%	73.3%	66.7%	66.7%
META IA	36	10	9	9	8
	70.6%	83.3%	60.0%	75.0%	66.7%
CLAUDE	41	11	12	9	9
	80.4%	91.7%	80.0%	75.0%	75.0%

Fuente: Elaboración propia con base en los resultados del piloto.

Los números reflejan la suma de los criterios aprobados en las diversas tablas de cumplimiento, junto con el porcentaje resultante en cada columna.

Por último, la Tabla 7 presenta una correlación entre el funcionamiento (puntuaje) y el cumplimiento para cada escenario (Acceso, Aplicación, Identidad e IPS) y cada plataforma (Check Point y Fortinet). Se indica también el total por fabricante y un total global por IA.

Tabla 7 Correlación Funcionamiento Vs Cumplimiento

CORRELACION FUNCIONAMIENTO VS CUMPLIMIENTO							
		ACCESO	APLICACION	IDENTIDAD	IPS	TOTAL FABRICANTE	
AI	Fabricante	Puntaje	Puntaje	Puntaje	Puntaje		TOTAL IA
CHATGPT	Checkpoint	4	4	2	3	13	20
	Fortinet	4	3	0	0	7	
GEMINI	Checkpoint	4	0	0	0	4	11
	Fortinet	4	0	0	3	7	
META IA	Checkpoint	3	0	0	0	3	10
	Fortinet	4	3	0	0	7	
CLAUDE	Checkpoint	4	0	0	0	4	14
	Fortinet	4	0	3	3	10	

Fuente: Elaboración propia con base en los resultados del piloto.

En esta tabla se distribuyen, para cada escenario, el valor “Puntaje” este se obtiene del producto entre los resultados de funcionalidad (Tabla 1) y el cumplimiento de cada escenario (Tabla 2 a Tabla 5). El “Total Fabricante” refleja la suma de cada escenario para Check Point o Fortinet por cada LLM, mientras que el “Total IA” indica la suma de puntajes a escala global para cada modelo.

4.1.3. Discusión

En esta sección se examinan de manera crítica los resultados presentados en el apartado anterior, con el propósito de explicar las diferencias observadas entre los distintos Modelos de Lenguaje (LLMs) y de destacar aquellos aspectos relevantes al objetivo general de la investigación. El análisis se centra en el desempeño de cada modelo (ChatGPT, Gemini, Meta IA y Claude) frente a los escenarios planteados (Control de Acceso, Control de Aplicaciones, Políticas basadas en Identidad y Prevención de Intrusiones), relacionando la efectividad operativa de las configuraciones con su adherencia a los lineamientos de seguridad.

4.1.3.1. Análisis de desempeño de cada LLM

ChatGPT. En los escenarios de Control de Acceso y Control de Aplicaciones, ChatGPT mostró un rendimiento destacable al proporcionar instrucciones operativas y relativamente alineadas con las buenas prácticas, incluyendo la definición de objetos y la habilitación de registros de eventos. En Políticas basadas en Identidad, pudo integrar referencias a servicios de directorio (Active Directory), aunque en ocasiones omitió configuraciones avanzadas vinculadas a la autenticación y restricción de perfiles de usuario. En Prevención de Intrusiones (IPS), generó secuencias de comandos útiles para activar firmas relacionadas con ataques como SQL Injection y XSS, si bien no siempre incluyó el nivel de granularidad necesario para la inspección de tráfico encriptado. Estos hallazgos sugieren que ChatGPT posee buena capacidad de adaptación al entorno del fabricante, especialmente cuando se le proporcionan prompts detallados y se segmenta adecuadamente la información, pero puede requerir supervisión experta en aspectos críticos de seguridad.

Gemini. El desempeño de Gemini en Control de Acceso fue sólido cuando las instrucciones eran concisas y no requerían una gran profundidad técnica, lo que se tradujo en la correcta

definición de reglas y objetos de firewall. Sin embargo, en Control de Aplicaciones y Políticas basadas en Identidad, evidenció fluctuaciones notables, posiblemente relacionadas con la forma en que procesó documentación extensa o más compleja (por ejemplo, integración de servicios de autenticación). En IPS, si bien pudo habilitar algunas firmas de detección, no siempre incluyó parámetros avanzados para el bloqueo automático de ataques. Su variabilidad entre escenarios apunta a que, al manejar grandes volúmenes de texto y configuraciones detalladas, Gemini tiende a presentar omisiones o inconsistencias, lo que pone de manifiesto la importancia de segmentar cuidadosamente la información que se le suministra.

Meta IA. Este modelo destacó en escenarios básicos de Control de Acceso y ofreció configuraciones iniciales para aplicaciones comunes en políticas de Control de Aplicaciones; no obstante, mostró dificultades para gestionar políticas basadas en Identidad, especialmente al relacionar directorios de usuarios con reglas específicas y grupos de seguridad. En IPS, pudo activar la detección de ataques conocidos, pero omitió con cierta frecuencia la definición de acciones de bloqueo o ajustes relacionados con la severidad de las firmas, lo que afectó su tasa de cumplimiento de los estándares. Estos comportamientos sugieren que Meta IA responde bien a peticiones directas y con menor complejidad técnica, pero puede quedarse corta cuando se le exige una configuración multidimensional que involucre parámetros avanzados o segmentación de usuarios y aplicaciones.

Claude. A lo largo de los cuatro escenarios, Claude evidenció un equilibrio razonable entre la funcionalidad de las configuraciones y la adecuación a las guías de seguridad. En Control de Acceso y Control de Aplicaciones, proporcionó respuestas que incluían la mayoría de los objetos y protocolos requeridos, así como la habilitación de registros. En Políticas basadas en Identidad, si bien no siempre detalló todas las fases de autenticación, sí demostró un manejo competente de la interacción con directorios de usuarios. En IPS, Claude generó reglas funcionales de protección contra ataques específicos, aunque no siempre cubrió la totalidad de los parámetros para la inspección de tráfico encriptado o la categorización avanzada de amenazas. Este patrón de resultados sugiere que, al recibir instrucciones claras y documentación adecuadamente estructurada, Claude puede producir configuraciones tanto operativas como relativamente completas, requiriendo ajustes puntuales para escenarios con altos requerimientos de seguridad.

4.1.3.2. Hallazgos por escenario de configuración

Control de Acceso. El planteamiento de reglas para permitir o denegar tráfico, definir direcciones, puertos y zonas se erigió en una tarea relativamente sencilla para la mayoría de los LLMs, al requerir pocos parámetros especializados. No obstante, se identificaron diferencias en el grado de exhaustividad con que los modelos abordaron opciones de registro y categorización de fuentes y destinos. Algunos sugerían la creación de grupos de objetos para mejorar la escalabilidad de la política, mientras que otros remitían a una configuración más puntual, definiendo únicamente la relación entre IPs y puertos sin ampliar la granularidad. Este contraste pone de relieve la influencia de la forma en que se estructura el prompt: si incluía referencias al uso de plantillas o reglas de logging, los LLMs eran más propensos a proponer directivas adicionales para asegurar trazabilidad y orden en la administración de accesos.

Control de Aplicaciones. El control a nivel de aplicación requirió a los modelos identificar correctamente protocolos y servicios, así como definir acciones específicas por categoría de tráfico. Aunque la mayoría de los LLMs propusieron reglas para aplicaciones populares (p. ej., aplicaciones de colaboración y P2P), solo algunos contemplaron políticas extensibles a otras herramientas corporativas potencialmente relevantes. Además, se observó que ciertos modelos descuidaban la configuración de inspección de tráfico cifrado —un aspecto vital para escudriñar amenazas encapsuladas—, lo cual indica la necesidad de que la documentación se presente de manera clara y segmentada para guiar la implementación de inspecciones TLS/SSL.

Políticas basadas en Identidad. La integración con directorios de usuarios (p. ej., Active Directory) implicó relacionar grupos específicos con perfiles de acceso, y este nivel de detalle se vio afectado por la precisión con que se facilitó la referencia al servicio de autenticación. Mientras que algunos LLMs incorporaron instrucciones genéricas para enlazar con la base de datos de usuarios, otros pasaron por alto detalles sobre la sincronización de credenciales o el modo de asignar privilegios de lectura/escritura en la consola del firewall. La variabilidad en estos aspectos recayó en la claridad de las guías o en la complejidad de las mismas, sobre todo cuando se requería un ajuste fino del tiempo de validez de tokens o la elaboración de listas blancas y negras en función de grupos de AD.

Prevención de Intrusiones (IPS). En la configuración de IPS, el foco se centró en la activación de firmas para ataques comunes (por ejemplo, SQL Injection) y la respuesta a esos eventos. Sin embargo, la profundidad con que los modelos abordaron la definición de acciones adicionales, como notificaciones vía correo electrónico, informes programados o despliegue en modo de simulación antes de su activación total, varió sustancialmente. Esto refuerza la noción de que ciertos LLMs responden bien a peticiones acotadas (activar IPS con firmas predefinidas) pero son menos precisos cuando deben detallar las fases de implementación (p. ej., modo detección vs. modo prevención, gestión de falsos positivos) o integrar reportes de incidentes en el flujo de seguridad global.

4.1.3.3. Relevancia para la Gestión de la Ciberseguridad

La verificación de un alto porcentaje de funcionamiento operacional (scripts ejecutables) no garantiza un cumplimiento total de los estándares de seguridad, lo que subraya la necesidad de validaciones adicionales. Por otro lado, la diversidad de resultados en cada escenario indica que la adopción de LLMs en procesos de configuración de firewall debe acompañarse de supervisión experta y de parámetros de control que aseguren la integridad y el apego a buenas prácticas.

El potencial de los LLMs para asistir a administradores inexpertos se ve reflejado en los casos en que se generaron configuraciones operativas y razonablemente alineadas con los lineamientos de fabricantes y estándares. Sin embargo, estas herramientas no constituyen un sustituto completo de la experiencia profesional, dado que persisten brechas técnicas y contextuales que pueden llevar a configuraciones incompletas o con riesgos encubiertos.

En síntesis, los resultados ponen de manifiesto tanto la promesa como las limitaciones de los LLMs en la generación de configuraciones de seguridad. ChatGPT y Claude mostraron un desempeño consistente en la mayoría de los escenarios, mientras que Gemini y Meta IA evidenciaron buenos resultados en entornos más simples pero fluctuaciones en configuraciones avanzadas. La adecuada segmentación de la documentación y la formulación cuidadosa de prompts se vislumbran como factores cruciales para optimizar la calidad de las respuestas.

5. Conclusiones y trabajo futuro

La presente investigación surgió a partir de la motivación de analizar el creciente uso de los Modelos de Lenguaje de Gran Escala (LLMs) en la ciberseguridad, especialmente en el contexto de la configuración de firewalls de próxima generación. Dada la accesibilidad y facilidad de uso de estas herramientas, se generó la pregunta sobre su fiabilidad y seguridad al compararlas con métodos convencionales.

el abordaje resultó pertinente, ya que permitió reflejar tanto el rendimiento de los LLM en un entorno operativo simulado como su grado de adherencia a lineamientos fundamentales. Las diversas tablas de evaluación y la estructura metodológica garantizaron que el análisis fuese sistemático y comparable, proveyendo así resultados objetivos que pueden extenderse a otros contextos de configuración.

A continuación, se exponen las conclusiones principales del estudio, se revisa el grado de cumplimiento de los objetivos propuestos y se plantean líneas de trabajo futuro que podrían enriquecer el campo de la ciberseguridad.

5.1.CONTRIBUCIONES Y RELACIÓN CON LOS OBJETIVOS PLANTEADOS

A lo largo del piloto experimental, se definieron objetivos concretos que guiaron tanto la elección de escenarios como la comparación de las configuraciones generadas por los LLMs. A continuación, se presentan los aportes clave en relación con dichos objetivos.

El estudio analizó el apoyo que pueden brindar los LLMs a profesionales con escasa experiencia, concluyendo que la mayoría de los modelos es capaz de producir configuraciones iniciales funcionales y, por ende, reducir la complejidad para quienes no cuentan con conocimientos profundos de administración de firewalls. No obstante, se identificaron omisiones en parámetros avanzados, lo que pone de manifiesto la necesidad de supervisión profesional.

La revisión de estas configuraciones frente a los CIS Benchmarks específicos para Fortinet y Check Point, así como al NIST SP 800-41, permitió cuantificar el grado de conformidad de las políticas propuestas. Si bien varios modelos exhibieron un nivel alto de cumplimiento en escenarios básicos, ninguno consiguió un acoplamiento completo sin intervenciones o ajustes adicionales.

Asimismo, se detectaron carencias recurrentes en la configuración de registros, la integración de autenticación en políticas basadas en identidad y la definición de acciones avanzadas en IPS. Estas limitaciones evidencian los riesgos de confiar en la salida de las IAs para entornos críticos de seguridad sin efectuar revisiones posteriores.

En cuanto a la efectividad de los LLMs para la creación de configuraciones seguras frente a administradores humanos con poca experiencia, se apreció que el tiempo de generación de reglas y guías puede disminuir. Sin embargo, todavía se requieren conocimientos fundamentales de ciberseguridad para identificar y corregir posibles incongruencias. En este sentido, los LLMs funcionan como un recurso de apoyo, aunque no sustituyen la intervención de un experto.

Por otra parte, en escenarios con requerimientos sencillos, se observó una clara reducción del tiempo de configuración, posibilitando la entrega acelerada de scripts de firewall listos para su validación. Aun así, la precisión resultó comprometida cuando las LLMs no recibieron una segmentación adecuada de la documentación del fabricante o cuando las necesidades de seguridad se volvían más complejas.

Finalmente, se constató que los LLMs pueden acortar la curva de aprendizaje, ya que ofrecen sugerencias y ejemplos concretos de configuración. De esta manera, facilitan la formación de nuevos administradores y sirven como punto de referencia para afianzar prácticas recomendadas y familiarizarse con ajustes clave en la protección de redes.

En su conjunto, estos hallazgos demuestran que los objetivos planteados se cumplieron de forma satisfactoria. Aunque ningún modelo está exento de fallas, la investigación arroja luz sobre el potencial de los LLMs y las precauciones que deben asumirse para su implementación efectiva y segura en entornos de ciberseguridad.

5.2. LÍNEAS DE TRABAJO FUTURO

A partir de los resultados y conclusiones, se ofrecen diversas oportunidades de investigación y optimización. En primer lugar, resulta clave profundizar en la alineación de seguridad, ya que la inspección de tráfico cifrado, la integración de directorios de usuarios y la definición de

parámetros IPS avanzados surgieron como puntos de mayor omisión. Para corregir estas deficiencias, podrían diseñarse prompts especializados que aborden a fondo dichos aspectos y fortalezcan la fiabilidad de los LLMs en entornos críticos.

Además, se propone la automatización de validaciones, mediante herramientas que comparen de forma automática las configuraciones generadas por LLMs con benchmarks de seguridad, de modo que destaquen errores u omisiones en tiempo real. Con ello se aliviaría la dependencia de personal experto, ya que las deficiencias podrían corregirse antes de la implementación en producción.

También se considera relevante ampliar los escenarios y fabricantes evaluados. Explorar entornos como Cisco, Palo Alto Networks o pfSense, y definir casos específicos para tecnologías emergentes (por ejemplo, la segmentación de contenedores o firewalls de microsegmentación), contribuiría a medir la robustez de las LLMs ante contextos más complejos o innovadores.

Por otra parte, la incorporación de técnicas de evaluación de riesgo mejoraría la visión sobre el impacto de las configuraciones generadas en la seguridad global de una organización. Se sugiere, por ejemplo, contrastar los ajustes con marcos como ISO 27001 o metodologías de análisis de amenazas, a fin de priorizar acciones que fortalezcan los niveles de protección.

Sumado a ello, se destaca la necesidad de colaboración multidisciplinaria y la adopción de estándares éticos. Dada la preocupación creciente por el uso responsable de la inteligencia artificial, sería conveniente alinear la creación de prompts y la evaluación de configuraciones con lineamientos que garanticen la transparencia y la protección de datos. La interacción entre profesionales técnicos, expertos jurídicos y asesores en ética contribuiría a lograr un empleo seguro y confiable de los LLMs en ciberseguridad.

Finalmente, se propone afianzar la formación continua a través de laboratorios virtuales que combinen configuraciones sugeridas por LLMs y prácticas manuales en un entorno controlado. Este enfoque permitiría a estudiantes y administradores inexpertos reforzar sus conocimientos teóricos, poner en práctica ajustes de seguridad reales y minimizar errores durante la fase de implementación.

5.3. NOTAS FINALES

En definitiva, la interacción con LLMs para la configuración de firewalls y otras soluciones de seguridad, si bien es prometedora, requiere de lineamientos claros, segmentación adecuada de la documentación y un control riguroso de los resultados. Los trabajos futuros propuestos ofrecen vías para superar las limitaciones halladas, solidificar la utilidad de estas herramientas en la práctica diaria y garantizar un uso responsable de la inteligencia artificial en entornos críticos de ciberseguridad.

El creciente uso de LLMs y su notable impacto en la ciberseguridad constituyen la base de la motivación y definición del problema de esta investigación, tal como se expone en la Sección 1.1. A partir de esta premisa, se evidencia la necesidad de establecer métodos de validación y supervisión para su implementación, a fin de prevenir riesgos operativos y de seguridad en infraestructuras críticas. Se espera que los aportes aquí presentados contribuyan a la implementación responsable y efectiva de herramientas basadas en IA, posibilitando una ciberseguridad más robusta e inclusiva para profesionales de todos los niveles de experiencia,

Referencias bibliográficas

Abdali, S., Anarfi, R., Barberan, C. J., & He, J. (2023). Securing large language models: Threats, vulnerabilities and responsible practices. arXiv. https://doi.org/10.48550/arXiv.2403.12503

Center for Internet Security. (2020). CIS Check Point Firewall Benchmark v1.1.0. Center for Internet Security. Recuperado de https://www.cisecurity.org

Center for Internet Security. (2024). CIS Fortigate 7.0.x Benchmark v1.3.0. Center for Internet Security. Recuperado de https://www.cisecurity.org

Deng, G., Liu, Y., Mayoral-Vilches, V., Liu, P., Li, Y., Xu, Y., Zhang, T., Liu, Y., Pinzger, M., & Rass, S. (2023). PENTESTGPT: Evaluating and harnessing large language models for automated penetration testing. arXiv. https://doi.org/10.48550/arXiv.2307.04657

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. arXiv. https://doi.org/10.48550/arXiv.2308.06782

He, Z., Li, Z., Yang, S., Qiao, A., Zhang, X., & Luo, X. (2023). Large language models for blockchain security: A systematic literature review. arXiv. https://doi.org/10.48550/arXiv.2403.14280

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Sun, R., Wang, Y., & Yang, Y. (2023). BEAVERTAILS: Towards improved safety alignment of LLM via a human-preference dataset. arXiv. https://doi.org/10.48550/arXiv.2307.04657

Sadiku, M. N. O., Fagbohungbe, O. I., & Musa, S. M. (2020). Artificial intelligence in cybersecurity. International Journal of Engineering Research and Advanced Technology, 6(5). https://doi.org/10.31695/IJERAT.2020.3612

Sandoval, G., Pearce, H., Nys, T., Karri, R., Garg, S., & Dolan-Gavitt, B. (2023). Lost at C: A User Study on the Security Implications of Large Language Model Code Assistants. 32nd USENIX Security Symposium. https://www.usenix.org/conference/usenixsecurity23/presentation/sandoval

Scarfone, K., & Hoffman, P. (2009). NIST Special Publication 800-41 Revision 1: Guidelines on Firewalls and Firewall Policy. National Institute of Standards and Technology. Recuperado de https://www.nist.gov

Tann, W., Liu, Y., Sim, J. H., Seah, C. M., & Chang, E.-C. (2023). Using Large Language Models for Cybersecurity: Capture-The-Flag Challenges and Certification Questions. National University of Singapore. https://arxiv.org/abs/2308.10443

Zhang, J., Bu, H., Wen, H., Liu, Y., Fei, H., Xi, R., Li, L., Yang, Y., Zhu, H., & Meng, D. (2025). When LLMs meet cybersecurity: A systematic literature review. Cybersecurity, 8(55). https://doi.org/10.1186/s42400-025-00361-w

Zhou, K., Hwang, J. D., Ren, X., & Sap, M. (2023). Relying on the unreliable: The impact of language models' reluctance to express uncertainty. arXiv. https://doi.org/10.48550/arXiv.2401.06730

Anexo A.

Anexo 1: Matriz de Resultados

Este anexo contiene la Matriz de Resultados, donde se documentan las configuraciones generadas por los modelos LLMs en los distintos escenarios evaluados. La matriz incluye:

- Las configuraciones propuestas por cada modelo para cada escenario.
- Los criterios de evaluación de cumplimiento, organizados según los estándares aplicables (CIS Benchmarks, NIST SP 800-41, entre otros).
- El análisis comparativo del grado de cumplimiento frente a los estándares establecidos.
- Los resultados consolidados de funcionalidad y cumplimiento, representando la correlación entre ambos factores.