COLEGIO DE CIENCIAS E INGENIERÍAS

INGENIERÍA EN CIENCIAS DE LA COMPUTACIÓN

Planificación para el Desarrollo del Proyecto Integrador

Tutor: Roberto Andrade

Autor: José Contreras

Quito – Ecuador

2025

1. Título del Proyecto:

   Orquestador de Agentes de IA para gestión de incidentes de ciberseguridad

2. Relevancia y Justificación:

   La creciente complejidad de las redes académicas y el aumento de ciber amenazas requieren soluciones avanzadas para gestionar incidentes de seguridad de manera eficiente. Este proyecto propone un orquestador de agentes de IA capaz de automatizar la detección, análisis, priorización y respuesta ante incidentes de ciberseguridad. El objetivo es optimizar tiempos de reacción y reducir riesgos mediante la coordinación de múltiples agentes especializados.

3. Objetivos

   a. Generales

      Diseñar e implementar un sistema orquestador de agentes de IA que gestione incidentes de ciberseguridad en redes, coordinando tareas de monitoreo, análisis y respuesta automatizada en una red, se proyecta reducir el MTTR ≥ 30 % y obtener una reducción de falsos positivos ≥ 20 % respecto de una línea base.

   b. Específicos

      • Desarrollar un agente de monitoreo para la detección temprana de amenazas.

      • Implementar un agente de análisis basado en RAG y LLMs para priorizar incidentes.

      • Crear un agente de respuesta que automatice notificaciones y acciones de mitigación.

      • Integrar todos los agentes y flujos mediante LangGraph y n8n.

      • Evaluar el rendimiento y efectividad del sistema.

4. Estado del Arte

En la gestión de incidentes de ciberseguridad, las soluciones tradicionales se han apoyado en dos pilares fundamentales: los sistemas de gestión de información y eventos de seguridad (SIEM) y las plataformas de orquestación, automatización y respuesta en seguridad (SOAR). Los SIEM tienen como propósito recolectar, normalizar y correlacionar grandes volúmenes de registros y eventos para ofrecer visibilidad centralizada sobre la infraestructura tecnológica; sin embargo, su aporte suele limitarse a la generación de alertas que requieren interpretación y acción humana, lo que prolonga los tiempos de respuesta y demanda de personal especializado (Islam, Babar & Nepal, 2019). Por su parte, las plataformas SOAR introducen flujos de automatización y *playbooks* predefinidos capaces de ejecutar respuestas semiautónomas, aunque su efectividad depende de integraciones costosas y rígidas con herramientas específicas, lo cual dificulta su adaptación a nuevas amenazas o a contextos empresariales cambiantes (Kremer et al., 2023).

Durante la última década, y con mayor énfasis desde 2020, la literatura ha mostrado un viraje hacia arquitecturas de orquestación multi-agente basadas en inteligencia artificial, que representan un paradigma más flexible y escalable que los enfoques anteriores. Estas arquitecturas articulan agentes especializados que cooperan bajo mecanismos de coordinación estructurados para cubrir funciones de monitoreo, análisis y respuesta (Castro et al., 2025). La evidencia empírica sugiere que la combinación de modelos de lenguaje de gran escala (LLMs), aprendizaje por refuerzo multi-agente (MARL) y metodologías de generación aumentada por recuperación (RAG) consigue mejoras significativas frente a SIEM y SOAR tradicionales: se reportan precisiones de detección superiores al 90%, tasas de falsos positivos entre 3,7% y 6,7% y correlaciones exitosas en la priorización de incidentes (MCC cercano a 0,99) en entornos de simulación (Blefari et al., 2025; Roelofs et al., 2024). De forma complementaria, los tiempos de reacción tienden a disminuir entre un 35% y un 45% cuando se despliegan agentes autónomos y marcos colaborativos tipo ChatOps, que permiten acelerar subtareas críticas de *triage* y contención (Lin et al., 2025; Brahmandam, 2025).

Este desplazamiento hacia enfoques multi-agente no obedece únicamente a razones técnicas, sino también a factores económicos y organizacionales. Mientras que las soluciones SIEM y SOAR suelen apoyarse en licencias propietarias, mantenimientos costosos e integraciones rígidas, los sistemas multi-agente se basan en arquitecturas modulares y portables, capaces de automatizar tareas repetitivas, reducir la dependencia de proveedores específicos y evolucionar de manera incremental al integrar o reemplazar agentes conforme cambian las tácticas de ataque (Song et al., 2024; Alshamrani, 2025). Esta flexibilidad refuerza la escalabilidad y resiliencia de los entornos de seguridad, constituyéndose en una alternativa sostenible frente a escenarios cambiantes.

En cuanto a aplicaciones prácticas, la literatura describe avances tanto en entornos simulados como en despliegues reales en SOC, servicios en la nube y sistemas críticos. La plataforma CyberRAG, orientada a la clasificación de ataques, alcanzó más del 94% de precisión y redujo los tiempos de *triage* en aproximadamente un 45% (Blefari et al., 2025). De manera similar, Audit-LLM, diseñado para la detección de amenazas internas mediante debate entre agentes, logró disminuir de forma significativa la tasa de falsos positivos frente a seis referentes convencionales (Song et al., 2024). Otros ejemplos son Triangle, que explora agentes negociadores para priorizar incidentes en entornos cloud, y CyGATE, que utiliza teoría de juegos para optimizar la planificación de parches en defensa adaptativa (Jiang et al., 2025).

No obstante, persisten limitaciones. Una parte considerable de los resultados proviene de simulaciones controladas que no siempre son extrapolables a redes empresariales complejas. Además, la rápida evolución de los LLM y de los marcos de orquestación puede volver obsoletas ciertas conclusiones en poco tiempo. Aunque ya existen experiencias en producción, todavía son escasas las comparaciones longitudinales y a gran escala con SIEM y SOAR que permitan estimar con precisión la magnitud de las mejoras operativas (Paduraru, Patilea & Stefanescu, 2025; Nyberg & Johnson, 2024).

Para abordar estos retos, la literatura recomienda el uso de métricas explícitas y reproducibles que permitan evaluar de forma objetiva el desempeño de los modelos y la eficiencia operativa del SOC. Dichas métricas se agrupan en cuatro familias: (i) métricas de modelo —precisión, *recall*, F1-score, AUC-ROC, tasa de falsos positivos (FPR), coeficiente de correlación de Matthews (MCC) y medidas de concordancia como τ de Kendall y ρ de Spearman—; (ii) métricas operativas y temporales —Mean Time to Detect (MTTD), Mean Time to Investigate (MTTI), Mean Time to Respond (MTTR) y Time to Act (TTA)—; (iii) métricas de analista y SOC —alertas por analista por día, tasa de falsos positivos por analista, tiempo medio de *triage*, porcentaje de alertas escaladas—; y (iv) métricas de madurez y negocio —cobertura de telemetría, posicionamiento en modelos de madurez como SOC-CMM, porcentaje de automatización y retorno sobre la inversión.

De manera transversal, diversos estudios coinciden en la necesidad de implementar esquemas de normalización de eventos tipo Common Event (CEC) que homogenicen campos críticos (*asset_id, rule_id, severity, confidence, IOCs, timestamps*) y preserven el evento crudo para auditoría y explicabilidad. Estos esquemas se han convertido en un patrón de facto en entornos SOC porque facilitan correlaciones confiables, consultas consistentes y la construcción de *dashboards* comparables a partir de datos provenientes de SIEM y logs de red, garantizando al mismo tiempo trazabilidad para revisiones posteriores y soporte a procesos de *post-incident review*.

En síntesis, el estado del arte muestra una transición desde infraestructuras centradas en SIEM y SOAR hacia arquitecturas multi-agente apoyadas en IA, acompañada de la consolidación de métricas estandarizadas y normalización de eventos como prácticas fundamentales para lograr sistemas de respuesta más eficientes, adaptables y auditables.

5. Metodología de Trabajo

La metodología propuesta se fundamenta en un ciclo de diseño de ingeniería alineado con los marcos de referencia internacionales más relevantes en gestión de incidentes, principalmente el NIST SP 800-61 y la ISO/IEC 27035-1:2023. Estos estándares proporcionan un marco estructurado que contempla preparación, detección, análisis, respuesta, recuperación y lecciones aprendidas, y sirven como guía para organizar tanto la definición de procesos como la evaluación de los resultados.

En una primera fase se llevará a cabo una revisión bibliográfica exhaustiva, con el objetivo de establecer las bases conceptuales del trabajo y de asegurar la correspondencia con las recomendaciones normativas. Esta revisión permitirá consolidar la taxonomía de métricas que se empleará en las etapas posteriores y establecer un marco de referencia que garantice la trazabilidad entre las actividades del proyecto y las buenas prácticas reconocidas internacionalmente.

La segunda fase se centrará en la caracterización del *threat landscape* bajo el contexto de gestión de incidentes. La principal fuente de información estará constituida por los registros recolectados en sistemas SIEM y logs de red. Estos datos permitirán identificar escenarios de amenaza de mayor relevancia para un entorno de operación de SOC, tales como ataques de denegación de servicio distribuido (DDoS), exfiltración de información, ransomware o amenazas internas. Para enriquecer esta etapa, se tomará como referencia el repositorio de playbooks de respuesta a incidentes de seguridad disponible en GitHub (MenakaGodakanda, 2024), que contiene plantillas y guías operativas para la gestión de diferentes tipos de ataque. A partir de este análisis se construirá un inventario de escenarios priorizados, mapeados a la taxonomía de MITRE ATT&CK y descritos con los indicadores mínimos de compromiso necesarios para su detección.

Posteriormente, se procederá a la definición de un esquema de normalización de eventos inspirado en las prácticas de *Common Event Canon (CEC)* descritas en la literatura. Este esquema estandarizará campos críticos tales como identificador de evento, marca temporal, fuente, regla de correlación, activo afectado, severidad, nivel de confianza y evento crudo preservado. Con esta estructura será posible integrar de forma consistente la información proveniente de diferentes fuentes de telemetría, manteniendo la trazabilidad necesaria para auditoría y *post-incident review*. Paralelamente se construirá un conjunto de referencia o *ground truth*, conformado a partir de incidentes verificados, tickets cerrados y evidencias confirmadas en el SIEM, complementado con casos controlados cuando se requiera asegurar cobertura de escenarios poco frecuentes. Este conjunto servirá como base para el entrenamiento y validación de los agentes de análisis.

La etapa de diseño e implementación contemplará la construcción modular de los flujos de detección, análisis y respuesta. En el flujo de detección, un programador temporizado activará la recolección de

registros y alertas desde el SIEM y otras fuentes de log, que serán normalizados bajo el esquema definido y almacenados en un repositorio de eventos. A partir de esta ingesta, el flujo de análisis realizará correlaciones mediante reglas heurísticas, consultará fuentes de reputación de indicadores de compromiso y aplicará técnicas basadas en modelos de lenguaje y recuperación aumentada (LLM/RAG) para clasificar incidentes, estimar su criticidad y establecer prioridades de atención. En paralelo, se generarán explicaciones en lenguaje natural que acompañarán cada decisión, con el fin de facilitar la comprensión y auditoría por parte de analistas humanos. Finalmente, el flujo de respuesta ejecutará de forma automatizada *playbooks* de contención y mitigación, incluyendo acciones como bloqueo de direcciones IP maliciosas, cuarentena de endpoints comprometidos, cierre de servicios críticos, apertura de tickets con evidencias adjuntas o rotación forzada de credenciales. Para incidentes de alta severidad se adoptará un esquema *human-in-the-loop*, que requerirá validación explícita antes de ejecutar acciones destructivas o irreversibles, en consonancia con los lineamientos de NIST e ISO.

La fase final corresponderá a la evaluación del sistema y la mejora continua. El desempeño será medido en comparación con una línea base definida por la operación manual con reglas tradicionales de SIEM. Se calcularán métricas de modelo como precisión, *recall*, F1-score, AUC-ROC, FPR y MCC; métricas operativas como MTTD, MTTI, MTTR y TTA; indicadores de carga humana como el número de alertas por analista y el tiempo medio de *triage*; y métricas organizacionales como cobertura de telemetría, porcentaje de automatización y madurez según modelos como SOC-CMM. Los resultados se registrarán con detalle de versiones de reglas, volúmenes de tráfico y condiciones de prueba para garantizar reproducibilidad. Asimismo, se implementarán *dashboards* con dos vistas complementarias: una operacional, orientada al monitoreo en tiempo casi real de volúmenes de alertas, tiempos de respuesta y estado de ejecución de *playbooks*; y otra ejecutiva, enfocada en tendencias mensuales de reducción de MTTR, evolución de tasas de falsos positivos, cobertura de escenarios y retorno sobre la inversión.

En conjunto, esta metodología busca demostrar que un sistema de orquestación basado en agentes inteligentes, sustentado en SIEM y logs de red como fuentes principales, es capaz de reducir significativamente los tiempos de respuesta y la tasa de falsos positivos, al mismo tiempo que proporciona trazabilidad, explicabilidad y alineamiento normativo. De esta forma se asegura la viabilidad técnica y operativa de un enfoque multi-agente, flexible y escalable para la gestión de incidentes de ciberseguridad en entornos contemporáneos de SOC.

A continuación, se incluye un diagrama preliminar para visualizar la arquitectura de los agentes y sus funciones:

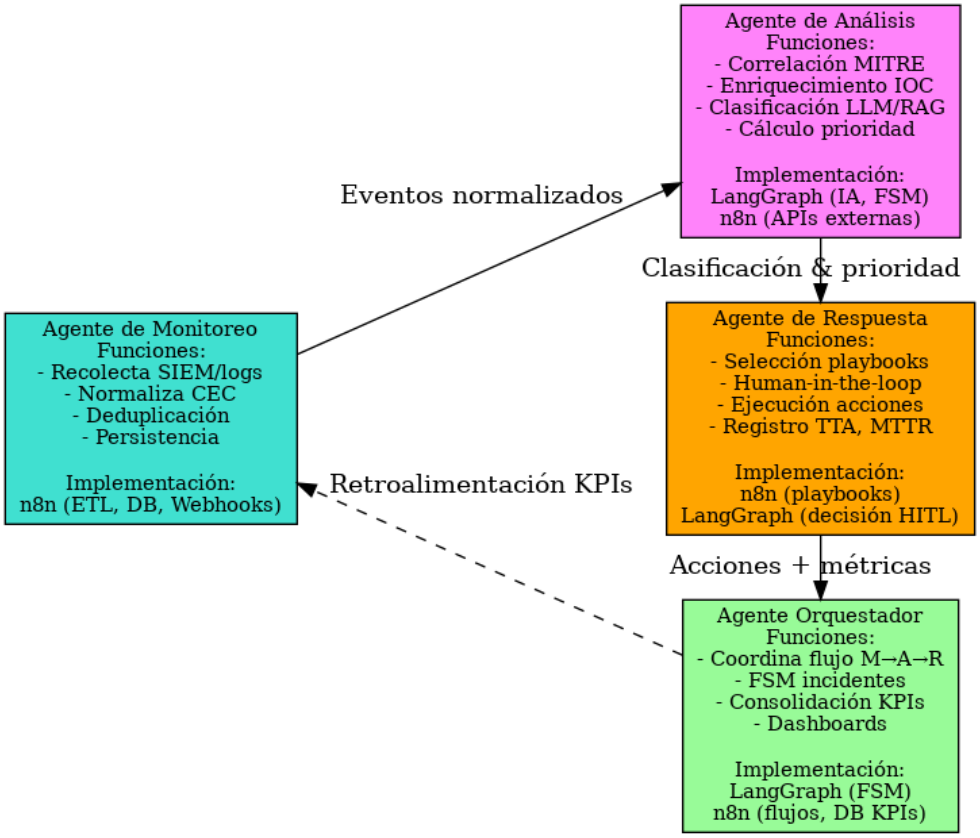

Figura 1. Diagrama preliminar de Arquitectura Agentes

El diagrama incluye la función de cada agente y su respectiva plataforma de implementación. También se puede observar la forma en la que se conectan a modo de pipeline de eventos. Es decir, el sentido dirigido por flechas representa también las etapas del ciclo del flujo a implementar.

6. Sumario de Contenidos

- Introducción
- Estado del Arte
- Descripción de la Propuesta
- Desarrollo del Prototipo
- Experimentos y Análisis de Resultados
- Conclusiones y Trabajo Futuro

7. Recursos

   a. Humanos

      Estudiante, tutor, profesores consultores, consultores externos, etc.

   b. Materiales

      Computadores personales, hardware de laboratorio

   c. Económicos

   Se contempla un presupuesto mínimo para licencias de software open-source y propietarios como LangGraph, Hostfinger, n8n entre otros según sea pertinente

8. Cronograma de Actividades

Actividades	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A1: Planificación final, baseline y diseño de arquitectura	X	X														
A2: Estado del arte ampliado; definición de escenarios			X	X												
A3: Agente de monitoreo + pipeline de eventos					X	X										
A4: Agente de análisis (RAG+clasificador) + evaluación preliminar							X	X								
A5: Agente de respuesta + 3 playbooks básicos									X	X						
A6: Orquestación con LangGraph y n8n; 2 playbooks adicionales											X	X				
A7: Campañas de prueba, análisis estadístico de KPIs, hardening													X	X		
A8: Preparación de defensa															X	
A9: Documento final y defensa																X

9. Entregables

• Documento de planificación (Semana 3)
• Entregable 1: Diseño de arquitectura y revisión del estado del arte (Semana 6)
• Entregable 2: Prototipo inicial con agentes básicos de monitoreo y análisis (Semana 10)
• Entregable 3: Prototipo avanzado de respuesta en conjunto con orquestación y resultado de pruebas (Semana 14)
• Documento final y defensa (Semana 16)

10. Referencias

- Alshamrani, A. (2025). *Federated hierarchical MARL for zero-shot cyber defense*. PLoS ONE. https://doi.org/10.1371/journal.pone.0329969
- Blefari, F., Cosentino, C., Pironti, F. A., Furfaro, A., & Marozzo, F. (2025). *CyberRAG: An agentic RAG cyber attack classification and reporting tool*. arXiv. https://doi.org/10.48550/arXiv.2507.02424
- Brahmandam, B. A. (2025). *AI driven ChatOps for DevSecOps: Automating security incident response*. International Journal of Multidisciplinary Research in Science, Engineering and Technology, 8(2), 85–96. https://doi.org/10.15680/ijmrset.2025.0802085
- Castro, S. R., Campbell, R., Lau, N., Villalobos, O., Duan, J., & Cardenas, A. A. (2025). *Large language models are autonomous cyber defenders*. Conference on Algebraic Informatics. https://doi.org/10.1109/CAI64502.2025.00195
- Islam, C., Babar, M. A., & Nepal, S. (2019). Automated interpretation and integration of security tools using semantic knowledge. In *Advanced Information Systems Engineering* (pp. 529–544). Springer. https://doi.org/10.1007/978-3-030-21290-2_32
- Jiang, Y., Oo, N., Meng, Q., Lin, L., Niyato, D., Xiong, Z., Lim, H. W., & Sikdar, B. (2025). *CyGATE: Game-theoretic cyber attack-defense engine for patch strategy optimization*. IEEE. https://doi.org/10.1109/CSR61664.2024.10679456
- Kremer, R., Wudali, P. N., Momiyama, S., Araki, J., Furukawa, J., Elovici, Y., & Shabtai, A. (2023). *IC-SECURE: Intelligent system for assisting security experts in generating playbooks for automated incident response*. arXiv. https://doi.org/10.48550/arXiv.2311.03825
- Lin, X., Zhang, J., Deng, G., Liu, T., Liu, X., Yang, C., Guo, Q., & Chen, R. (2025). *IRCopilot: Automated incident response with large language models*. arXiv. https://doi.org/10.48550/arXiv.2505.11901
- Nyberg, J., & Johnson, P. (2024). *Structural generalization in autonomous cyber incident response with message-passing neural networks and reinforcement learning*. Computer Science Symposium in Russia. https://doi.org/10.1109/CSR61664.2024.10679456
- Paduraru, C., Patilea, C., & Stefanescu, A. (2025). *CyberGuardian 2: Integrating LLMs and agentic AI assistants for securing distributed networks*. International Conference on Evaluation of Novel Approaches to Software Engineering. https://doi.org/10.5220/0013406000003928
- Roelofs, T.-M., Bárbaro, E., Pekarskikh, S., Orzechowska, K., Kwapien, M., Tyrlik, J., Smadu, D., van Eeten, M., & Zhauniarovich, Y. (2024). *Finding harmony in the noise: Blending security orchestration with large language models for effective incident triage*. arXiv. https://doi.org/10.48550/arXiv.2407.12345

11. Revisión y firma del tutor del proyecto

Yo, Roberto Andrade, profesor de la carrera de Ingeniería en Ciencias de la Computación, hago constar que he revisado y, por lo tanto, apruebo el documento de planificación del proyecto titulado “Orquestador de Agentes de IA para Gestión de Incidentes de Ciberseguridad” propuesto por el estudiante Jose Contreras. Por otra parte, me comprometo a proporcionar al estudiante el soporte necesario y oportuno para el buen desarrollo del proyecto antes mencionado.

Firmado electrónicamente por:
ROBERTO OMAR
ANDRADE PAREDES
Validar únicamente con FirmaEC

Fdo: Roberto Andrade

Quito, 3 de Octubre de 2025