

Diseños de redes de LLM de IA generativa frente a metodología tradicional humana

Generative AI LLM network designs vs. traditional human methodology

Fausto Raúl Orozco Lara¹ (fausto.orozcol@ug.edu.ec) (<https://orcid.org/0000-0003-4872-3702>)

Yturalde Villagómez Juan Carlos² (juan.yturaldev@ug.edu.ec) (<https://orcid.org/0000-0001-9948-177X>)

Nicolas Fernando Ahtty Arteaga³ (nicolas.ahttya@ug.edu.ec) (<https://orcid.org/0009-0000-6411-3803>)

Resumen

Este artículo surge como producto final del trabajo de titulación que tiene como tema “Evaluación comparativa de rendimiento y convergencia en arquitecturas de red generadas mediante inteligencia artificial generativa frente a diseños en base a modelos de ingeniería tradicional”, cuyo objetivo es determinar si la inteligencia artificial posee capacidades técnicas y cognitivas suficientes para diseñar redes de forma autónoma, mediante simulación controlada. Se definieron diez casos de red heterogéneos, documentados públicamente, para generar diseños experimentales con tres modelos de lenguaje de gran escala, contrastados contra un grupo de control de diseño tradicional. Los cuarenta diseños resultantes se evaluaron mediante ochenta simulaciones en el simulador de eventos discretos ns-3, en escenarios de operación normal y de resiliencia ante fallas. El análisis combinó estadística inferencial no paramétrica (Wilcoxon, Friedman) sobre throughput, latencia y pérdida de paquetes, con un análisis cualitativo de consistencia interna de los diseños entregados. Los resultados muestran que la inteligencia artificial replica con fidelidad estadísticamente equivalente la arquitectura lógica del diseño de referencia, pero subestima de forma significativa y con efecto grande la capacidad de enlace requerida. Se identificaron además diferencias de fiabilidad entre modelos: un modelo no presentó incidencias de consistencia interna, mientras que los otros dos entregaron diseños

¹ Universidad de Guayaquil, Ecuador

² Universidad de Guayaquil, Ecuador

³ Universidad de Guayaquil, Ecuador

incompletos o con afirmaciones contradichas por su propia topología. Se concluye una capacidad parcial y heterogénea de la inteligencia artificial generativa para el diseño autónomo de redes, concentrada principalmente en el dimensionamiento de capacidad como su limitación más crítica.

Abstract

This research is the final product of the degree thesis titled "Comparative Evaluation of Performance and Convergence in Network Architectures Generated through Generative Artificial Intelligence versus Traditional Engineering-Based Designs," whose objective is to determine whether artificial intelligence possesses sufficient technical and cognitive capabilities to autonomously design networks, through controlled simulation. Ten heterogeneous, publicly documented network cases were defined to generate experimental designs using three large language models, contrasted against a traditional-design control group. The resulting forty designs were evaluated through eighty simulations in the discrete-event simulator ns-3, under normal operation and fault-resilience scenarios. The analysis combined non-parametric inferential statistics (Wilcoxon, Friedman) on throughput, latency, and packet loss with a qualitative analysis of the internal consistency of the delivered designs. The results show that artificial intelligence replicates the logical architecture of the reference design with statistically equivalent fidelity, but significantly and with a large effect underestimates the required link capacity. Reliability differences among models were also identified: one model presented no internal consistency issues, while the other two delivered incomplete designs or designs containing statements contradicted by their own topology. It is concluded that generative artificial intelligence has a partial and heterogeneous capacity for autonomous network design, concentrated primarily around capacity dimensioning as its most critical limitation.

Palabras clave: inteligencia artificial, diseño de redes, simulación de redes, modelos de lenguaje de gran escala, resiliencia de red

Keywords: artificial intelligence, network design, network simulation, large language models, network resilience

Introducción

Este estudio analiza que el fenómeno de la incorporación de la inteligencia artificial (IA) en procesos de ingeniería ha dejado de ser una tendencia incipiente para convertirse en un componente habitual de la práctica técnica contemporánea. Allison et al. (2022) sostienen que la IA constituye ya una herramienta transversal para el diseño en ingeniería, capaz de intervenir en fases que tradicionalmente dependían exclusivamente del criterio experto humano, mientras que Wankhede et al. (2024) destacan su papel como facilitador de la toma de decisiones técnicas en entornos de incertidumbre, siempre que se mantenga como apoyo y no como sustituto absoluto del juicio profesional. En el ámbito hispanoamericano, Rodríguez Acasio (2024) documenta un patrón similar de adopción creciente de la IA dentro de las disciplinas de la ingeniería, advirtiendo que la evaluación empírica de sus resultados continúa siendo escasa frente a la velocidad de su adopción práctica.

Dentro de las disciplinas de ingeniería, el diseño y la configuración de redes de datos se ha convertido en un dominio particularmente activo para la aplicación de modelos de lenguaje de gran escala (LLM, por sus siglas en inglés). Wang et al. (2024), en un estudio desarrollado en el KTH Royal Institute of Technology en colaboración con NVIDIA y Red Hat, evaluaron la capacidad de distintos LLM (GPT-3.5, GPT-4, GPT-4-Turbo y CodeLlama) para traducir requerimientos de red expresados en lenguaje natural hacia especificaciones formales, llamadas de función y configuraciones de bajo nivel, encontrando que los modelos más avanzados logran una precisión cercana al 100% en tareas simples y estructuradas, pero que su desempeño se degrada en escenarios complejos, donde persisten alucinaciones y errores dependientes del tamaño del contexto que exigen verificación y supervisión humana. Un patrón similar de asistencia orientada al análisis, más que a la generación, se observa en Shan et al. (2025), quienes desarrollaron RCA Copilot, un sistema basado en LLM diseñado para transformar datos de telemetría y monitoreo de red en información procesable para el diagnóstico de causa raíz de incidentes. Al igual que GeNet, este tipo de herramienta opera sobre una infraestructura ya desplegada y datos operativos existentes, con el propósito de asistir al ingeniero en la interpretación y el diagnóstico, y no de producir una arquitectura de red completa a partir de

requerimientos iniciales. Este tipo de LLMs orientados al análisis, la configuración o el diagnóstico de redes ya existentes, contrasta con el enfoque del presente estudio, centrado exclusivamente en la capacidad de modelos de lenguaje de propósito general para generar diseños de topología de red completos y autónomos desde cero, sin infraestructura ni datos operativos previos como insumo, y se descartan como modelos seleccionable para la investigación. Adicionalmente, Qiu et al. (2024) desarrollaron un optimizador combinatorio basado en un LLM, llamado LMCO, para la planificación de puntos de acceso inalámbrico, demostrando que dicho optimizador supera al algoritmo de referencia Ant Colony Optimization particularmente a medida que aumenta la complejidad del problema, lo cual evidencia que la superioridad de la IA generativa frente a métodos tradicionales no es uniforme, sino que depende del tipo específico de tarea de diseño evaluada.

Esta heterogeneidad de resultados se confirma en investigaciones centradas específicamente en la interpretación de topologías. Ifland et al. (2025), en un trabajo desarrollado en la Universidad Ben-Gurion del Néguev en colaboración con Fujitsu, presentaron GeNet, un copiloto multimodal basado en LLM capaz de interpretar imágenes de topología de red y generar actualizaciones de configuración a partir de intenciones expresadas en lenguaje natural, validado sobre escenarios adaptados de ejercicios de certificación Cisco. Sus resultados muestran una interpretación visual precisa de la topología, pero también identifican un margen de esfuerzo de corrección no trivial antes de que las soluciones generadas sean equivalentes al diseño de referencia. En el dominio específico de redes de internet de las cosas (IoT), Alipio y Bures (2025) revisaron el estado del arte del uso de LLM para mejorar protocolos de red, automatizar decisiones y fortalecer la seguridad, concluyendo que, si bien estos modelos ya intervienen en configuración de red y detección de amenazas, persisten retos relacionados con el alto costo computacional, los riesgos de seguridad y la adaptación lenta en redes dinámicas.

La limitación más recurrente entre estos antecedentes es la dificultad de los modelos generativos para satisfacer restricciones técnicas estrictas. Du et al. (2024) documentan que los modelos de difusión generativa aplicados a la optimización de redes ofrecen ventajas notables frente al aprendizaje por refuerzo profundo tradicional en la exploración de espacios de solución

complejos, pero advierten que su naturaleza probabilística dificulta garantizar el cumplimiento de restricciones duras. En la misma dirección, Yang et al. (2026) presentan una revisión crítica de la IA generativa aplicada a la optimización de redes, señalando como limitaciones intrínsecas la dificultad para satisfacer restricciones estrictas, la comprensión conceptual limitada de los modelos y la naturaleza inherentemente probabilística de sus salidas, factores que condicionan directamente la fiabilidad de cualquier diseño técnico que dichos modelos produzcan de forma autónoma.

Este patrón no es exclusivo del dominio de redes. Karamthulla et al. (2023) compararon distintos enfoques de IA (aprendizaje automático, algoritmos evolutivos y aprendizaje por refuerzo profundo) para la asignación dinámica de recursos en infraestructura de nube, encontrando diferencias cuantificables de desempeño entre métodos según la métrica evaluada, un patrón de comparación metodológica análogo al que la ingeniería de redes requiere para validar la sustitución del criterio humano por soluciones automatizadas. La ausencia de este tipo de validación comparativa constituye, precisamente, el vacío que motiva la presente investigación, que es la adopción de topologías generadas por IA en entornos de producción suele producirse sin que exista evidencia experimental suficiente sobre si dichos diseños alcanzan niveles de rendimiento, confiabilidad y viabilidad técnica comparables a los del diseño tradicional basado en el criterio de un ingeniero humano, particularmente en infraestructuras de misión crítica donde errores de dimensionamiento u omisiones lógicas no detectadas pueden derivar en interrupciones de servicio, degradación sostenida del rendimiento e incumplimiento de acuerdos de nivel de servicio.

El objetivo general de la presente investigación fue determinar si la inteligencia artificial posee las capacidades técnicas y cognitivas suficientes para la creación de diseños de redes de forma autónoma, mediante un análisis de desempeño en un entorno de simulación controlado. Para ello se plantearon cuatro objetivos específicos: definir los parámetros de requerimientos y tablas de rendimiento aplicables a topologías de redes en estructuras complejas; generar topologías de red basadas en IA con configuraciones de equipos conforme a dichos parámetros; evaluar el comportamiento de las métricas definidas bajo condiciones de carga de tráfico concurrente en

simulación; y divulgar los hallazgos del análisis comparativo mediante la postulación de un artículo científico en una revista indexada. Se planteó como hipótesis de investigación que las topologías de red diseñadas mediante inteligencia artificial presentan un nivel de confiabilidad técnica inferior al de aquellas diseñadas por ingeniería humana, bajo condiciones de simulación controladas.

Por sus objetivos es una investigación aplicada; por su alcance, experimental, al incorporar la manipulación deliberada del origen del diseño como variable independiente; y por su lugar de desarrollo, de laboratorio, al ejecutarse íntegramente en un entorno de simulación controlado y reproducible.

Materiales y métodos

La investigación se desarrolló bajo el enfoque metodológico de Investigación en Ciencias del Diseño (Design Science Research), orientado a la resolución de un problema técnico mediante la construcción y evaluación sistemática de un artefacto experimental, organizado en cuatro fases: identificación del problema, diseño del artefacto, evaluación y comunicación. Por su modalidad, el estudio combina un componente cuantitativo, centrado en la recolección y el análisis estadístico de indicadores objetivos de rendimiento, con un componente cualitativo complementario.

El marco muestral consta de diez arquitecturas de red heterogéneas, seleccionadas mediante muestreo intencional por criterios de diversidad arquitectónica, procedencia confiable y compatibilidad con simulación. Ocho de estas redes corresponden a infraestructuras de producción real, documentadas y verificadas por consorcios académicos e institucionales, tomadas del repositorio The Internet Topology Zoo, y a través de plataformas oficiales de las instituciones dueñas de cada red. Las dos redes restantes corresponden a arquitecturas en base a diseños guía de fabricantes, red tipo campus jerárquico, red SD-WAN, centro de datos Spine-Leaf y una topología hub-and-spoke en la nube, se consideran como diseños elaborados a partir de guías oficiales de CISCO Catalyst y Microsoft Azure. La Tabla 1 resume los datos importantes de cada caso de red.

Tabla 1. Casos de red seleccionados

Caso	Nodos /Enlaces	Topología lógica	Enrutamiento
GARR	50 / 49	Spine-Leaf de 3 capas	IS-IS / OSPF
Internet2	59 / 70	Jerárquica de dos capas con malla parcial	OSPF/IS-IS con BGP externo
GEANT	50 / 56	Malla parcial, sin anillo perimetral completo	OSPF/IS-IS con SRv6/SR-MPLS
ESNET	40 / 43	Malla redundante con doble conexión a hubs de tránsito en puntos críticos	OSPF/IS-IS con Segment Routing
OXFORD	12 / 21	Doble estrella redundante activo-activo (2 núcleos Core)	OSPF de área única con BGP en el borde hacia JANET
CEDIA	45 / 59	Triángulo núcleo entre Quito-Guayaquil-Cuenca	OSPFv3
SDWAN	20 / 60	Multi regional con malla completa	OMP propietario
CAMPUS	22 / 39	Jerárquica de 3 capas con doble enlace activo-activo	No especificado
DATA CENTER	26 / 60	Fabric Clos de 2 niveles, malla completa entre niveles	EBGP de un salto
HUB AND SPOKE	10 / 9	Hub-and-Spoke de un solo nivel, 2 Hubs redundantes interconectados	sin protocolo dinámico real

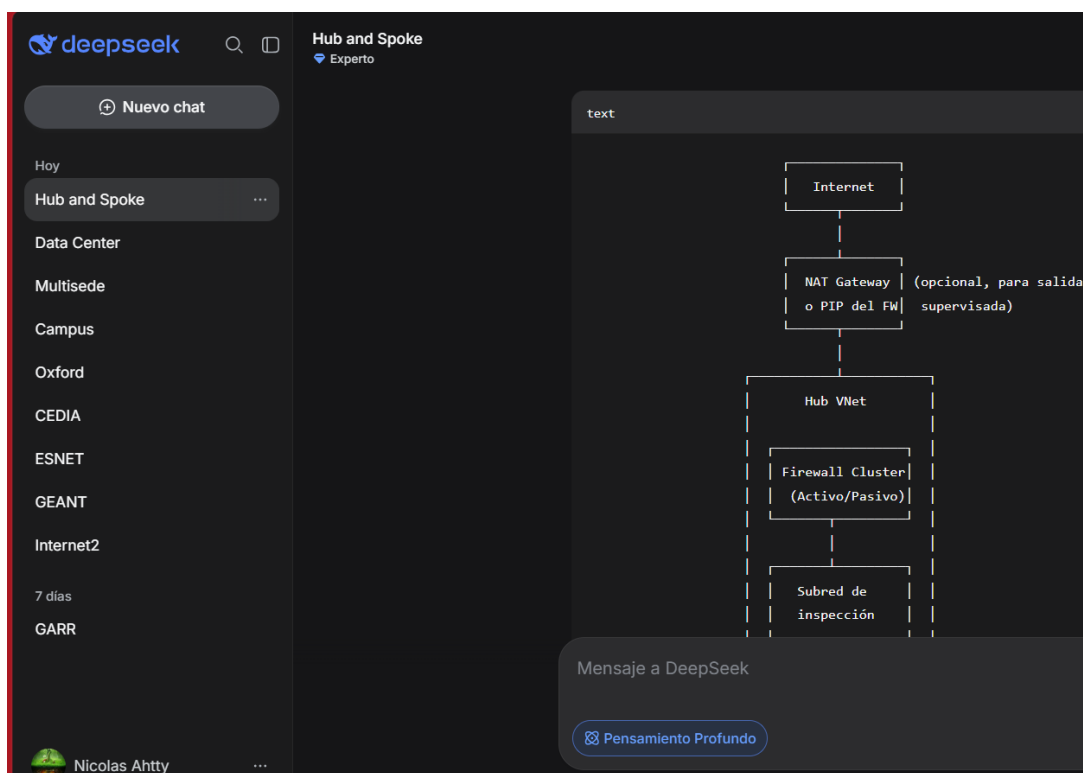
Fuente: Autor de la investigación (2026).

Para cada uno de los diez casos se obtuvo un diseño de origen tradicional, tomado directamente de la arquitectura documentada, y tres diseños de origen generativo, producidos de manera independiente por tres LLMs, los cuales son Claude Sonnet 4.6, Deepseek-V4 y OpenAI o3, estos tres modelos se han seleccionado por su mayor capacidad en tareas de análisis matemáticos y comprensión de textos extensos. Deepseek es utilizada en la plataforma oficial, Sonnet

igualmente por la plataforma Claude de la empresa Anthropic, y OpenAI es un modelo disponible en la plataforma Chat GPT-5.5 perteneciente a la empresa OpenAI.

La Figura 1 muestra la interacción hecha con Deepseek, mostrando un chat distinto para cada caso de red; este LLM no genera archivos, pero si detalla el diseño en la misma plataforma.

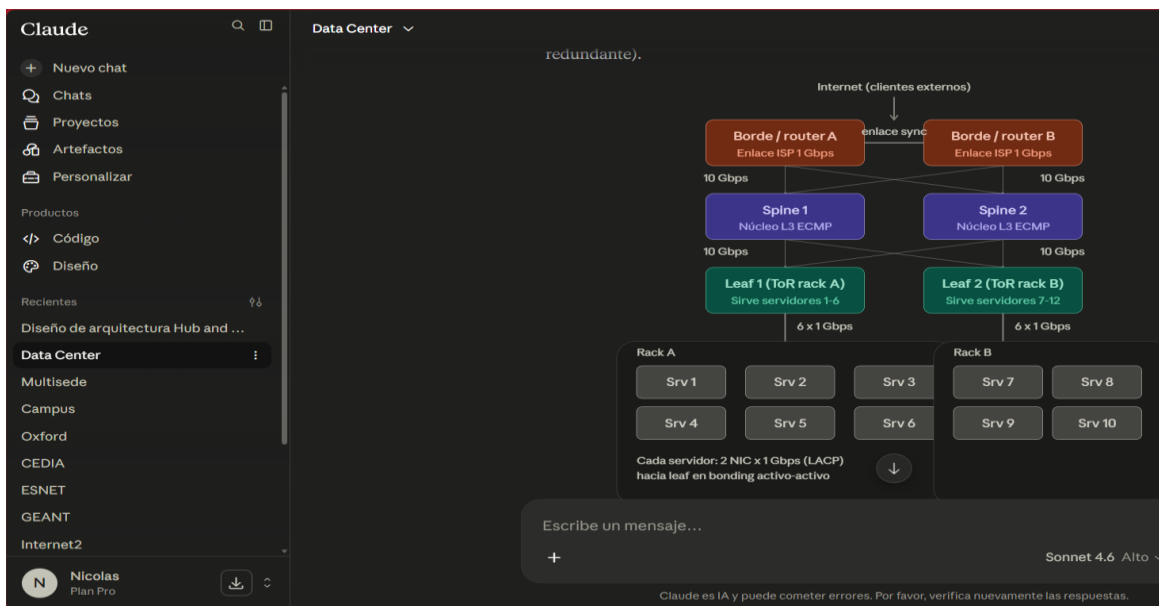
Figura 1. Interacción con Deepseek



Fuente: Autores de la investigación (2026).

La Figura 2 muestra la interacción hecha con Sonnet, mostrando un chat distinto para cada caso de red; este LLM sí genera archivos, obteniendo uno por cada caso de red.

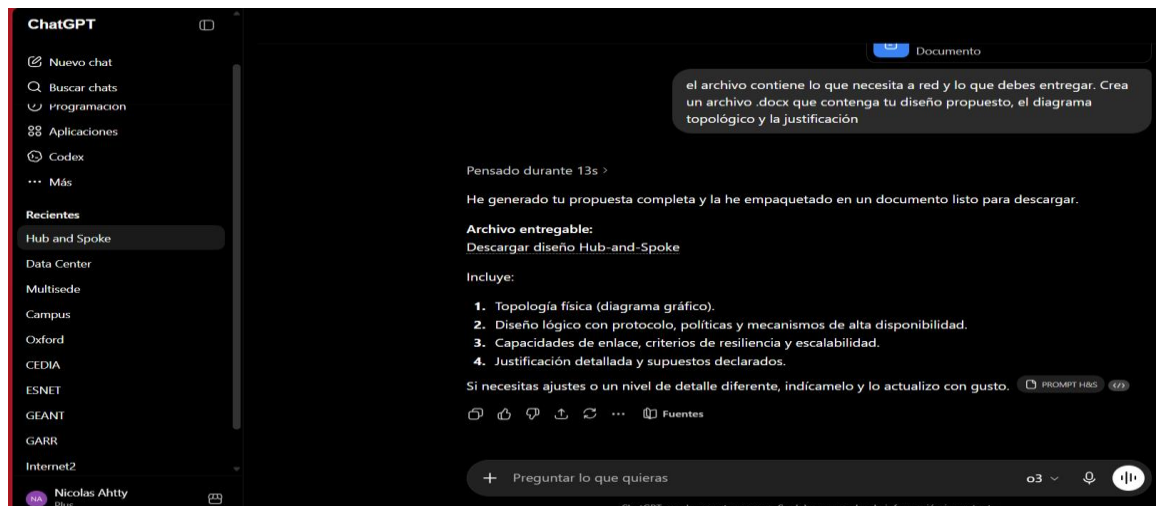
Figura 2. Interacción con Sonnet



Fuente: Autores de la investigación (2026).

La Figura 3 muestra la interacción hecha con Sonnet, mostrando un chat distinto para cada caso de red; este LLM también genera archivos, obteniendo uno por cada caso de red.

Figura 3. Interacción con OpenAI



Fuente: Autores de la investigación (2026).

Se les solicitó crear sus propios diseños mediante la creación de diez guiones de interacción, usando un protocolo estandarizado de guiónprompt engineering basado en el principio de caja negra, que restringía a cada modelo a los mismos datos de entrada (nodos, funciones, tecnología de transporte y objetivo de servicio) sin inducir decisiones de diseño. Este procedimiento produjo un total de cuarenta unidades de estudio, correspondientes a un grupo de control de diez diseños tradicionales y un grupo experimental de treinta diseños generados por IA.

La evaluación experimental se ejecutó en el simulador de eventos discretos ns-3. La selección de esta herramienta se sustenta en la evidencia sistemática reportada por Campanile et al. (2020), quienes, mediante una revisión sistemática de literatura, documentan la adopción amplia de ns-3 en múltiples dominios de investigación en redes, incluyendo redes locales, redes estructuradas de gran escala y backbones de centros de datos, y su validez científica frente a otras herramientas de simulación. Se descartó el uso de emuladores en tiempo real como alternativa, dado que Ottens et al. (2025) demuestran que el modo de emulación en tiempo real de ns-3 impone limitaciones considerables sobre el throughput alcanzable y la complejidad topológica soportada, en comparación con la simulación de eventos discretos sin restricciones de tiempo real, lo que habría comprometido la validez comparativa entre los cuarenta diseños evaluados. La pertinencia de ns-3 para la evaluación comparativa de diseños de red mediante métricas de rendimiento estandarizadas ha sido documentada además por Malnar y Jevtić (2020), quienes desarrollaron un framework de evaluación de redes vehiculares basado precisamente en throughput, pérdida de paquetes, sobrecarga, retardo extremo a extremo y jitter, y por Baumgartner et al. (2021), quienes validaron el uso de ns-3 para modelar tecnologías de acceso de nueva generación mediante las mismas métricas de retardo y throughput empleadas en este estudio. La extensibilidad de ns-3 para incorporar componentes de inteligencia artificial en la simulación de red ha sido evidenciada por Ekaireb et al. (2022) mediante ns3-fl, un simulador que integra ns-3 con marcos de aprendizaje federado bajo condiciones de red realistas, y su vigencia como entorno de validación continúa siendo respaldada por trabajos recientes presentados en el propio foro oficial de la

comunidad ns-3, como el de Shen et al. (2024) sobre validación de retardo en operación multienlace.

Se mantuvieron constantes, para los cuarenta diseños, el motor de simulación, el criterio de medición de cada métrica mediante el módulo FlowMonitor, el patrón y volumen de tráfico generado según la categoría funcional de cada red, y el instante y mecanismo de la falla inducida, de modo que cualquier diferencia observada en las métricas fuese atribuible al criterio de diseño evaluado y no a variaciones del proceso de simulación. Se ejecutaron dos escenarios por diseño, un escenario induciendo tráfico normal y otro de resiliencia ante una falla de enlace inducida al segundo 60 de una simulación de 120 segundos, lo que produjo un total de ochenta observaciones experimentales. Las métricas registradas fueron latencia media (ms), jitter medio (ms), throughput (Mbps) y porcentaje de pérdida de paquetes, calculadas de forma automatizada a partir de los archivos XML exportados por FlowMonitor tras cada simulación.

Dado que la unidad de análisis estadístico corresponde a los diez casos de red evaluados de forma pareada, y no a las ochenta simulaciones individuales, el tamaño muestral no permite aplicar pruebas paramétricas que exigen un mínimo de treinta observaciones independientes. Por ello, siguiendo el marco de selección de pruebas estadísticas para comparación de modelos descrito por Rainio et al. (2024), se emplearon exclusivamente pruebas no paramétricas que son la prueba de rangos con signo de Wilcoxon, para comparar el grupo de control frente a cada modelo de IA por métrica, y la prueba de Friedman, para comparar los tres modelos de IA entre sí, usando los umbrales convencionales de Cohen. En la Tabla 2 se puede observar los resultados del test de Wilcoxon por métrica analizada, comparando cada LLM con la metodología tradicional.

Tabla 2. Resultados del test de Wilcoxon

Métrica	Modelo	W	p- valor	Sig.	R	Efecto	Comparación
Throughput tráfico normal	Sonnet	8.0	0.0977	ns	0.629	grande	IA PEOR
	o3	2.0	0.0117	*	0.822	grande	IA PEOR
	DeepSeek	5.0	0.0391	*	0.725	grande	IA PEOR

Throughput resiliencia	Sonnet	9.0	0.1289	ns	0.596	grande	IA PEOR
	o3	2.0	0.0117	*	0.822	grande	IA PEOR
Latencia tráfico normal	DeepSeek	3.0	0.0195	*	0.790	grande	IA PEOR
	Sonnet	24.0	0.7695	ns	0.113	pequeño	IA PEOR
	o3	23.0	0.6953	ns	0.145	pequeño	IA MEJOR
	DeepSeek	15.0	0.2324	ns	0.403	mediano	IA MEJOR
Packet loss resiliencia	Sonnet	11.0	0.1055	ns	0.532	grande	IA MEJOR
	o3	16.0	0.2754	ns	0.371	mediano	IA MEJOR
	DeepSeek	9.0	0.0645	ns	0.596	grande	IA MEJOR

Fuente: Autores de la investigación (2026).

La significancia indica si la diferencia se debe a un factor directo del diseño, o no, siendo “ns” un valor no significativo y un asterisco (*) representa una significancia real de diferencia de diseños.

El análisis cualitativo, por su parte, se desarrolló mediante análisis de contenido dirigido sobre los documentos de diseño entregados por cada modelo, verificando su consistencia interna entre la descripción textual, la tabla de enlaces declarada y la topología reconstruida mediante teoría de grafos, así como su cumplimiento de los requisitos explícitos del protocolo de prompt engineering aplicado. La Tabla 3 muestra la comparativa entre los distintos LLM con sus respectivos pares de diseño tradicional por cada métrica evaluada.

Tabla 3. Comparativa por métrica de estudio

Variable	Metodología con mejor desempeño	Valor de probabilidad	Magnitud De rango
Throughput	Tradicional	$p > 0,7$	Grande ($r > 0,72$)
Latencia	Equivalentes	$p > 0,23$	Pequeño a mediano
Packet loss	Tendencia: IA	todos $p > 0,06$	Mediano a grande ($r = 0,37 - 0,60$)
Downtime estimado	Tendencia: IA	indicador indirecto	6/10 casos favorables

Fuente: Autor de la investigación (2026).

Resultados

El análisis descriptivo, calculado sobre los diez casos de red del marco muestral, muestra que el grupo de control obtuvo un throughput medio de 2,12 Mbps en el escenario de tráfico normal, frente a medias de 1,30 Mbps para el modelo de mejor desempeño entre las tres inteligencias artificiales evaluadas y de 0,97 Mbps para el de menor desempeño. Esta ventaja del diseño tradicional se replicó en el escenario de resiliencia y se mantuvo en siete de los diez casos individuales, con una brecha especialmente marcada en las topologías de tipo SD-WAN y campus jerárquico, donde la reducción de throughput de los diseños de IA superó el 89% frente al diseño de referencia. El coeficiente de variación del throughput fue sistemáticamente más alto en los tres modelos de IA que en el grupo de control, lo que indica que, además de operar con una capacidad promedio menor, los diseños generados por IA resultan proporcionalmente más heterogéneos entre casos de red.

En la métrica de latencia bajo tráfico normal, en cambio, no se observó una dirección sistemática de diferencia en la que algunos casos los modelos de IA presentaron menor latencia que el diseño tradicional y en otros, mayor. El coeficiente de correlación de Pearson entre el ordenamiento relativo de los diez casos del grupo de control y el de cada modelo de IA fue moderado a alto para esta métrica (entre 0,55 y 0,58), lo que indica que los tres modelos conservan razonablemente el ordenamiento de latencia del diseño de referencia entre casos, a diferencia de lo observado en throughput, donde dicha correlación fue prácticamente nula (entre 0,03 y 0,14).

En pérdida de paquetes durante el escenario de resiliencia, el patrón favoreció mayoritariamente al grupo experimental, puesto que, en siete de los diez casos, al menos dos de los tres modelos de IA registraron menor pérdida que el diseño tradicional, con reducciones medianas de entre 39% y 80% según el modelo. El caso más favorable correspondió al modelo con fortaleza declarada en estructuración de grafos, con una mediana de pérdida de 0,13%, frente a 1,66% del grupo de control.

La prueba de Wilcoxon de rangos con signo confirmó que la ventaja del diseño tradicional en throughput es estadísticamente significativa y de efecto grande frente a dos de los tres modelos de

IA evaluados, tanto en el escenario normal (r entre 0,72 y 0,82; $p < 0,05$) como en el de resiliencia, mientras que frente al tercer modelo la diferencia mostró la misma dirección y una magnitud de efecto igualmente grande ($r \approx 0,60$ a $0,63$), aunque sin alcanzar significancia formal debido a su mayor variabilidad entre casos. En latencia bajo tráfico normal, ningún modelo de IA presentó diferencias estadísticamente significativas frente al diseño tradicional ($p > 0,23$ en los tres casos), con tamaños de efecto pequeños a medianos, resultado que respalda que los tres modelos comprenden de forma adecuada la arquitectura lógica de conectividad de las redes evaluadas. En pérdida de paquetes durante resiliencia, ningún modelo alcanzó significancia formal ($p > 0,06$), aunque los tres mostraron efecto mediano a grande en dirección favorable a la inteligencia artificial.

La prueba de Friedman, aplicada para comparar los tres modelos de IA entre sí, mostró diferencias estadísticamente significativas en throughput tanto en el escenario normal ($p = 0,018$) como en el de resiliencia ($p = 0,027$), con un modelo consistentemente superior a los otros dos, mientras que en latencia normal y en pérdida de paquetes de resiliencia los tres modelos resultaron estadísticamente equivalentes entre sí ($p = 0,703$ y $p = 0,378$, respectivamente). El ranking promedio calculado sobre las cuarenta comparaciones posibles (diez casos por cuatro métricas) ubicó al diseño tradicional en el primer lugar global, con una posición promedio de 2,20, seguido de cerca por el mejor modelo de IA con 2,27, mientras que los otros dos modelos de IA obtuvieron posiciones promedio de 2,70 y 2,83.

Debido a que el tráfico simulado empleó una tasa constante (CBR), el jitter resultó mayoritariamente nulo o idéntico entre las cuatro metodologías de diseño en la mayoría de los casos, por lo que se excluyó del análisis inferencial y se reportó únicamente de forma descriptiva. De manera análoga, la instrumentación del tiempo de detección de falla y del tiempo de reconvergencia no constituyó una medición válida de la velocidad de recuperación de cada diseño, el primero correspondió a una constante de configuración del experimento y el segundo estuvo correlacionado con la densidad de flujos de cada caso más que con el diseño evaluado, por lo que ambas variables se excluyeron del análisis cuantitativo y la dimensión de convergencia se

operacionalizó, en su lugar, a través de la magnitud de la pérdida de paquetes durante el escenario de resiliencia.

El análisis cualitativo de consistencia interna, aplicado sobre los treinta documentos de diseño entregados por los modelos de IA, identificó una distribución marcadamente desigual de incidencias entre modelos. Uno de los tres modelos no presentó ninguna incidencia de diseño incompetente ni de incongruencia entre representaciones en los diez casos evaluados. El segundo modelo dejó nodos obligatorios sin conectar en la primera entrega en tres de los diez casos, requiriendo una solicitud explícita de continuación, y presentó seis incidencias de incongruencia entre el texto descriptivo, la tabla de enlaces y la topología verificada mediante teoría de grafos, cuatro de ellas de severidad crítica por declarar explícitamente una propiedad de redundancia o resiliencia que la verificación topológica desmintió para una parte sustancial de la red, hasta el 80% de los nodos en uno de los casos. El tercer modelo presentó tres incidencias de esta misma categoría, concentradas en los casos de mayor escala y complejidad estructural del marco muestral.

Discusión

Los hallazgos de esta investigación son consistentes con el patrón de capacidad parcial documentado en la literatura reciente sobre IA aplicada al diseño de redes. La comprensión adecuada de la arquitectura lógica de conectividad, reflejada en la equivalencia estadística de latencia y en la correlación moderada a alta con el ordenamiento del diseño tradicional, coincide con la evidencia de Qiu et al. (2024), quienes encontraron que un optimizador combinatorio basado en LLM supera al algoritmo de referencia Ant Colony Optimization en tareas de planificación de puntos de acceso, particularmente cuando aumenta la complejidad del problema. Sin embargo, la limitación sistemática y estadísticamente significativa observada en el dimensionamiento de capacidad de enlace no se explica por una falla de comprensión topológica, sino que es coherente con la advertencia de Yang et al. (2026) sobre la dificultad intrínseca de los modelos generativos para satisfacer restricciones técnicas duras, y con la observación de Du et al. (2024) de que la naturaleza probabilística de estos modelos dificulta garantizar el cumplimiento

estricto de parámetros cuantitativos, a diferencia de decisiones estructurales de tipo más cualitativo, como la elección de una topología en anillo, en malla o jerárquica.

El análisis cualitativo de este estudio aporta evidencia adicional a la limitación documentada por Wang et al. (2024), quienes advirtieron que incluso los modelos de mejor desempeño en tareas de configuración de red permanecen propensos a alucinaciones y requieren verificación humana en escenarios complejos. Las incidencias de incongruencia entre representaciones registradas en el presente estudio, en las cuales un modelo declara explícitamente una propiedad de redundancia que su propia topología no cumple, constituyen una manifestación concreta y verificable de ese mismo fenómeno, aplicada específicamente al dominio de diseño de topologías. Este hallazgo también dialoga con la metodología de evaluación propuesta por Ifland et al. (2025) para sistemas LLM multimodales aplicados a topología de red, cuyas métricas de fidelidad de solución y esfuerzo de corrección asumen, precisamente, que la salida de un modelo generativo requiere de un proceso de verificación posterior antes de considerarse equivalente al diseño de referencia esperado.

La heterogeneidad de fiabilidad encontrada entre los tres modelos evaluados, uno sin incidencias documentadas frente a otros dos con múltiples incongruencias críticas, resulta relevante para la discusión planteada por Alipio y Bures (2025) sobre los retos pendientes en la aplicación de LLM a redes de internet de las cosas, entre ellos la adaptación heterogénea de distintos modelos ante tareas de configuración y seguridad. El presente estudio sugiere que dicha heterogeneidad no solo se manifiesta en el desempeño cuantitativo, sino también en la fiabilidad con la que cada modelo comunica su propio diseño, una dimensión cognitiva distinta de la dimensión técnica de dimensionamiento y que resulta especialmente crítica para un usuario sin formación en ingeniería de redes, que no dispondría de las herramientas de verificación empleadas en este estudio para detectar que una propiedad declarada por el modelo es falsa.

La tendencia, no significativa, pero consistente, hacia una menor pérdida de paquetes de los diseños de IA durante el escenario de resiliencia matiza la interpretación de una inferioridad generalizada de la IA generativa. Este patrón es compatible con el hallazgo de Karamthulla et al. (2023), quienes documentaron que distintos enfoques de IA pueden superar a métodos de

referencia en dimensiones específicas de desempeño, en su caso, tiempo de respuesta y utilización de CPU en infraestructura de nube, sin que ello implique una superioridad integral del enfoque automatizado. El patrón hallado también resulta metodológicamente análogo al reportado por Shahid et al. (2024) en su comparación de protocolos de enrutamiento EIGRP, OSPF y BGP bajo escenarios de conmutación por falla, en la cual un protocolo superó a otro en tiempo de convergencia y pérdida de paquetes sin ser categóricamente mejor en todas las métricas evaluadas, evidenciando que las comparaciones de desempeño en redes rara vez producen una superioridad absoluta de un enfoque sobre otro, sino ventajas concentradas en dimensiones específicas.

La validez de esta comparación descansa, en buena medida, en la calidad del marco muestral empleado. El uso de topologías reales documentadas por consorcios académicos y el repositorio Internet Topology Zoo, permitió contrastar los diseños de IA frente a arquitecturas efectivamente desplegadas en producción, en lugar de frente a topologías sintéticas o simplificadas, lo que fortalece la validez externa de los hallazgos frente a estudios que emplean escenarios de prueba de menor escala. La adecuación metodológica de emplear pruebas no paramétricas de Wilcoxon y Friedman ante un tamaño muestral reducido, siguiendo el criterio de selección de pruebas estadísticas descrito por Rainio et al. (2024), resultó determinante para poder distinguir con rigor las diferencias sistemáticas, como la de throughput- de aquellas atribuibles al azar -como la de latencia y pérdida de paquetes, en lugar de forzar una prueba paramétrica no aplicable al número de casos disponibles.

Finalmente, el patrón encontrado converge con la discusión más amplia sobre el rol de la IA como herramienta de apoyo y no como sustituto pleno del criterio humano, planteada por Allison et al. (2022) y por Wankhede et al. (2024) para la ingeniería en general, y particularizada por Rodríguez Acasio (2024) para el contexto de adopción tecnológica hispanoamericano. Este mismo patrón de evaluación comparativa entre un sistema basado en IA y un estándar de referencia tradicional, mediante un diseño experimental estructurado, es análogo al empleado por Giannakopoulos et al. (2023) para comparar el desempeño de ChatGPT, Google Bard y Microsoft Bing Chat frente a la evidencia científica tradicional en un dominio completamente distinto, la

odontología basada en evidencia, cuyos resultados también mostraron que ningún modelo de lenguaje igualó de forma consistente el nivel de la referencia tradicional en todas las dimensiones evaluadas, reforzando la idea de que la capacidad parcial y heterogénea de los LLM frente a estándares humanos de referencia no es un hallazgo aislado del dominio de redes, sino un patrón que se repite en dominios técnicos diversos.

Conclusiones

La inteligencia artificial posee capacidades técnicas y cognitivas parciales para la creación autónoma de diseños de redes: reproduce de forma estadísticamente equivalente al diseño humano la lógica de conectividad de una topología, mostrando una correlación de ordenamiento de latencia entre 0,55 y 0,58, frente a $p > 0,23$ sin diferencias significativas en los tres modelos. Pero no alcanza su nivel de desempeño en el dimensionamiento de la capacidad de enlace, con un throughput medio de 2,12 Mbps del diseño tradicional frente a 1,30 Mbps y 0,97 Mbps de los modelos de IA, con r entre 0,72 y 0,82 y $p < 0,05$ frente a dos de los tres modelos, que constituye la limitación más crítica y estadísticamente robusta identificada en el estudio. Esta capacidad parcial no es, además, uniforme entre modelos, dado que la fiabilidad con la que cada uno comunica su propio diseño varió sustancialmente, con incidencias de diseño incompleto y de incongruencia interna concentradas en dos de los tres modelos evaluados, con nodos sin conectar en 3 de 10 casos y 6 incongruencias, 4 de severidad crítica, en un modelo; 3 incidencias en el otro; ninguna incidencia en el tercero.

Sí es factible definir un conjunto de parámetros de requerimientos y métricas de rendimiento aplicables de forma homogénea a topologías de red de distinta complejidad, lo cual permitió sostener una comparación técnica válida entre metodologías de diseño, en que la inteligencia artificial es capaz de generar topologías de red funcionales con configuraciones coherentes con los parámetros establecidos, aunque no siempre de forma completa ni consistente en su primera entrega; y que, bajo condiciones de tráfico concurrente, las topologías generadas por IA presentan un comportamiento estadísticamente inferior en capacidad de red (Friedman $p = 0,018$ en tráfico normal y $p = 0,027$ en resiliencia, con posiciones promedio de 2,20 para el diseño tradicional frente a 2,27, 2,70 y 2,83 para los modelos de IA), equivalente en conectividad lógica, y con una

tendencia, no significativa, favorable a la IA en pérdida de paquetes ante fallas (reducciones medianas de entre 39% y 80% según el modelo, $p > 0,06$).

En relación con la hipótesis de investigación, la evidencia obtenida ofrece una confirmación parcial y no absoluta, ya que la confiabilidad técnica de los diseños generados por IA resultó efectivamente inferior a la del diseño humano de forma estadísticamente significativa en la dimensión de capacidad de red, pero no en las dimensiones de conectividad lógica ni de resiliencia ante fallas, donde ambas metodologías se comportaron como estadísticamente equivalentes, con un valor probabilístico (p) superior a 0,23 en latencia y superior a 0,06 en pérdida de paquetes. La hipótesis se sostiene, por tanto, de manera focalizada en la dimensión de dimensionamiento de capacidad, y no como una inferioridad generalizada de la inteligencia artificial frente al criterio humano en todas las dimensiones técnicas evaluadas.

Estos hallazgos sugieren que la incorporación de inteligencia artificial generativa al diseño de redes de datos debe concebirse, en el estado actual de la tecnología, como una herramienta de apoyo al criterio del ingeniero de redes y no como un sustituto autónomo de dicho criterio, particularmente en lo referente al dimensionamiento de capacidad de enlace y a la verificación de las propiedades de resiliencia que el propio modelo declara en su diseño, dado que estas últimas no siempre resultaron consistentes con la topología efectivamente entregada, teniendo más del 80% de los nodos afectados en uno de los casos documentados.

Referencias

- Alipio, M., & Bures, M. (2025). The role of large language models in designing reliable networks for Internet of Things: A short review of most recent developments. *IEEE Access*, 13, 168527-168545. <https://doi.org/10.1109/ACCESS.2025.3599281>
- Allison, J. T., Cardin, M. A., McComb, C., Ren, M. Y., Selva, D., Tucker, C., Vermillion, C., Yukish, M., & Zhao, Y. F. (Eds.). (2022). Artificial intelligence and engineering design. *Journal of Mechanical Design*, 144(2), Article 020301. <https://doi.org/10.1115/1.4053111>

- Baumgartner, M., Juhár, J., & Papaj, J. (2021). Simulation of 5G and LTE-A access technologies via network simulator ns-3. En 2021 44th International Conference on Telecommunications and Signal Processing (TSP) (pp. 77-80). IEEE. <https://doi.org/10.1109/TSP52935.2021.9522659>
- Campanile, L., Gribaudo, M., Iacono, M., Marulli, F., & Mastroianni, M. (2020). Computer network simulation with ns-3: A systematic literature review. Electronics, 9(2), Article 272. <https://doi.org/10.3390/electronics9020272>
- Du, H., Zhang, R., Liu, Y., Wang, J., Lin, Y., Li, Z., Niyato, D., Kang, J., Xiong, Z., Cui, S., Ai, B., Zhou, H., & Kim, D. I. (2024). Enhancing deep reinforcement learning: A tutorial on generative diffusion models in network optimization. IEEE Communications Surveys & Tutorials, 26(4), 2611-2646. <https://doi.org/10.1109/COMST.2024.3400011>
- Ekaireb, E., Yu, X., Ergun, K., Zhao, Q., Lee, K., Huzaifa, M., & Rosing, T. (2022). ns3-fl: Simulating federated learning with ns-3. En Proceedings of the 2022 Workshop on ns-3 (WNS3 '22) (pp. 97-104). ACM. <https://doi.org/10.1145/3532577.3532591>
- Giannakopoulos, K., Kavadella, A., Aaqel Salim, A., Stamatopoulos, V., & Kaklamanos, E. G. (2023). Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: Comparative mixed methods study. Journal of Medical Internet Research, 25, Article e51580. <https://doi.org/10.2196/51580>
- Ifland, B., Krief, R., Zilberman, A., Duani, E., Ohana, M., Murillo, A., Manor, O., Lavi, O., Hikichi, K., Shabtai, A., Elovici, Y., & Puzis, R. (2025). GeNet: A multimodal LLM-based co-pilot for network topology and configuration. En Proceedings of the 2025 IEEE 45th International Conference on Distributed Computing Systems Workshops (ICDCSW) (pp. 117-122). IEEE. <https://doi.org/10.1109/ICDCSW63273.2025.00026>

- Karamthulla, M. J., Malaiyappan, J. N. A., & Tillu, R. (2023). Optimizing resource allocation in cloud infrastructure through AI automation: A comparative study. *Journal of Knowledge Learning and Science Technology*, 2(2), 1-10. <https://doi.org/10.60087/jklst.vol2.n2.p326>
- Malnar, M., & Jevtić, N. (2020). A framework for performance evaluation of VANETs using NS-3 simulator. *Promet – Traffic&Transportation*, 32(2), 255-268. <https://doi.org/10.7307/ptt.v32i2.3227>
- Ottens, M., Deutschmann, J., Hielscher, K.-S., & German, R. (2025). Performance evaluation of ns-3 real-time emulation. *IEEE Access*, 13. <https://doi.org/10.1109/ACCESS.2025.3555478>
- Qiu, K., Bakirtzis, S., Wassell, I., Song, H., Zhang, J., & Wang, K. (2024). Large language model-based wireless network design. *IEEE Wireless Communications Letters*, 13(12), 3340-3344. <https://doi.org/10.1109/LWC.2024.3468676>
- Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14, Article 6086. <https://doi.org/10.1038/s41598-024-56706-x>
- Rodríguez Acasio, F. J. (2024). Alcance de la inteligencia artificial en la ingeniería. *Ingenium et Potentia. Revista Electrónica Multidisciplinaria de Ciencias Básicas, Ingeniería y Arquitectura*, 6(10), 1-3. <https://doi.org/10.35381/i.p.v6i10.3873>
- Shahid, K., Ahmad, S. N., & Rizvi, S. T. H. (2024). Optimizing network performance: A comparative analysis of EIGRP, OSPF, and BGP in IPv6-based load-sharing and link-failover systems. *Future Internet*, 16(9), Article 339. <https://doi.org/10.3390/fi16090339>
- Shan, A., Kaur, J., Singh, R., Banka, T., Yavatkar, R., & Sridhar, T. (2025). *RCA Copilot: Transforming network data into actionable insights via large language models*. En *2025 IEEE International Conference on Communications (ICC)* (pp. 1566–1571). IEEE. <https://doi.org/10.48550/arXiv.2507.03224>
- Shen, M., Zhang, J., Yin, H., Roy, S., & Gao, Y. (2024). Delay in multi-link operation in ns-3: Validation and impact of traffic splitting. En *Proceedings of the 2024 Workshop on ns-3 (WNS3 '24)* (pp. 19-26). ACM. <https://doi.org/10.1145/3659111.3659116>

- Wang, C., Scazzariello, M., Farshin, A., Ferlin, S., Kostić, D., & Chiesa, M. (2024). NetConfEval: Can LLMs facilitate network configuration? Proceedings of the ACM on Networking, 2(CoNEXT2), 1-25. <https://doi.org/10.1145/3676851>
- Wankhede, V. A., Agrawal, R., Kumar, A., Luthra, S., Pamucar, D., & Stević, Ž. (2024). Artificial intelligence an enabler for sustainable engineering decision-making in uncertain environment: A review and future propositions. Journal of Global Operations and Strategic Sourcing, 17(2), 384-401. <https://doi.org/10.1108/JGOSS-06-2022-0057>
- Yang, B., Liang, R., Li, W., Wang, H., Cao, X., Yu, Z., Lasaulce, S., Debbah, M., Alouini, M.-S., Poor, H. V., & Yuen, C. (2026). Frontiers of generative AI for network optimization: Theories, limits, and visions. IEEE Communications Surveys & Tutorials, 28, 4555-4580. <https://doi.org/10.1109/COMST.2026.3656241>