

Evaluación de arquitecturas RAG descentralizadas para la reducción de alucinaciones y mejora de la trazabilidad en entornos corporativos simulados

Evaluation of decentralized RAG architectures for hallucination reduction and improved traceability in simulated corporate environments.

Diego Parraga Menchaca^{1*} 

¹ Universidad Mayor de San Andrés, Facultad de Ciencias Puras y Naturales, Edificio de la carrera de Informática, La Paz, Bolivia

*Autor de correspondencia: parragadiego@gmail.com

Recibido: 05/05/2026 Aceptado: 03/09/2026 Publicado: 13/10/2026

DOI: <https://doi.org/10.37518/2663-6360x2026v9n1p1>

RESUMEN: La incorporación de modelos de lenguaje de gran escala en entornos organizacionales requiere mecanismos que permitan mejorar la recuperación de información y reducir la generación de respuestas no sustentadas en fuentes verificables. El objetivo del estudio fue evaluar una arquitectura de generación aumentada por recuperación (Retrieval-Augmented Generation, RAG) descentralizada orientada a mejorar la trazabilidad de la información y reducir las alucinaciones en entornos corporativos simulados. Se implementó un entorno experimental basado en un corpus sintético de 10000 documentos corporativos distribuidos en repositorios vectoriales federados. La arquitectura propuesta integró cuatro componentes principales: ingesta distribuida, almacenamiento vectorial federado, enrutamiento semántico y mecanismos de validación de respuestas mediante guardrails. Se evaluaron escenarios comparativos entre arquitecturas centralizadas y descentralizadas utilizando métricas de tiempo hasta el primer token (TTFT), tasa de alucinación, fidelidad de respuesta mediante RAGAS y validación humana. Los resultados evidenciaron una reducción del TTFT desde 3,2 s hasta 0,4 s en consultas complejas y una disminución de la tasa de alucinación desde 32,0 % hasta 1,9 % en el entorno experimental evaluado. Asimismo, la validación humana mostró una concordancia sustancial en la identificación de respuestas alucinadas (Kappa de Fleiss = 0,82). Se concluye que las arquitecturas RAG descentralizadas contribuyeron a mejorar la eficiencia de recuperación y la trazabilidad de las respuestas generadas en entornos corporativos simulados. No obstante, se requieren estudios adicionales para validar estos resultados en escenarios organizacionales reales.

Palabras clave: auditoría informática, bases de datos vectoriales, ciberseguridad, inteligencia artificial.

ABSTRACT: The adoption of Large Language Models in organizational environments requires mechanisms that improve information retrieval while reducing unsupported or hallucinated responses. The objective of this study was to evaluate a decentralized Retrieval-Augmented Generation (RAG) architecture designed to enhance information traceability and reduce hallucinations in simulated corporate environments. An experimental environment was implemented using a synthetic corpus of 10,000 corporate documents distributed across federated vector repositories. The proposed architecture integrated four main components: distributed data ingestion, federated vector storage, semantic routing, and response validation mechanisms through guardrails. Comparative scenarios between centralized and decentralized architectures were evaluated using Time-To-First-Token (TTFT), hallucination rate, response faithfulness measured with RAGAS, and human validation. The results showed a reduction in TTFT from 3.2 s to 0.4 s for complex queries and a decrease in hallucination rate from 32.0 % to 1.9 % within the evaluated experimental environment. Human assessment also demonstrated substantial agreement in hallucination detection (Fleiss' Kappa = 0.82). The findings suggest that decentralized RAG architectures can improve retrieval efficiency and response traceability in simulated corporate environments. However, further validation in real organizational settings is required before generalizing these results.

Keywords: artificial intelligence, cybersecurity, vector databases, IT auditing.

1. INTRODUCCIÓN

La transformación digital ha incrementado significativamente el volumen y la diversidad de datos generados por las organizaciones. Como consecuencia, la información corporativa suele encontrarse distribuida entre repositorios locales, sistemas empresariales, plataformas en la nube y múltiples fuentes documentales, dificultando su integración y explotación eficiente para la toma de decisiones (Davenport & Bean, 2023; Gao et al., 2024).

En los últimos años, los Grandes Modelos de Lenguaje (Large Language Models, LLMs) han demostrado capacidades sobresalientes para la generación de contenido y la asistencia en procesos de análisis de infor-

mación. Sin embargo, estos modelos presentan limitaciones importantes cuando requieren acceder a conocimiento especializado o actualizado que no forma parte de sus parámetros de entrenamiento. Asimismo, pueden generar respuestas incorrectas o no verificables, fenómeno conocido como alucinación (hallucination) (Ji et al., 2023; Mialon et al., 2023).

Con el propósito de superar estas limitaciones, Lewis et al. (2020) propusieron la arquitectura de Generación Aumentada por Recuperación (Retrieval-Augmented Generation, RAG), la cual combina mecanismos de recuperación documental con modelos generativos. Este enfoque permite que el modelo consulte información externa antes de construir una respuesta, mejorando la precisión y el acceso a conocimiento actualizado (Lewis

et al., 2020; Gao et al., 2024). No obstante, las implementaciones tradicionales de RAG suelen depender de repositorios centralizados que pueden generar cuellos de botella cuando el volumen documental aumenta o las consultas se distribuyen entre diferentes dominios organizacionales (Wang et al., 2024).

La literatura reciente propone arquitecturas descentralizadas basadas en repositorios vectoriales federados y técnicas de enrutamiento semántico para mejorar la eficiencia de recuperación. En estos entornos, los documentos permanecen distribuidos en múltiples nodos y el sistema selecciona únicamente las fuentes más relevantes para cada consulta. Este enfoque puede reducir el espacio de búsqueda, disminuir la latencia y mejorar la pertinencia del contexto suministrado al modelo generativo (Wang et al., 2024; Karpukhin et al., 2020).

Paralelamente, la incorporación de mecanismos de validación conocidos como guardrails permite verificar la coherencia entre las respuestas generadas y las fuentes recuperadas. Estas estrategias favorecen la trazabilidad de la información utilizada por el sistema y contribuyen a reducir la generación de respuestas fuera de contexto o inconsistentes con la evidencia documental disponible (Es et al., 2024; Ji et al., 2023).

A pesar de los avances reportados, aún existe limitada evidencia experimental acerca del impacto combinado del almacenamiento vectorial federado, el enrutamiento semántico distribuido y los mecanismos de validación sobre métricas objetivas de desempeño, particularmente en escenarios que simulan entornos corporativos de gran escala. Esta situación evidencia la necesidad de evaluar arquitecturas descentralizadas capaces de mejorar simultáneamente la eficiencia de recuperación y la confiabilidad de las respuestas generadas.

En este contexto, el objetivo de la presente investigación fue evaluar una arquitectura RAG descentralizada orientada a la reducción de alucinaciones y la mejora de la trazabilidad de respuestas en entornos corporativos simulados. Para ello, se comparó una arquitectura centralizada con una arquitectura descentralizada utilizando indicadores de tiempo hasta el primer token (Time-To-First-Token, TTFT) y tasa de alucinación.

Con base en este objetivo se formularon las siguientes hipótesis:

- H1: Las arquitecturas RAG descentralizadas presentan menores tiempos de respuesta (TTFT) que las arquitecturas RAG centralizadas bajo condiciones equivalentes de consulta.
- H2: Las arquitecturas RAG descentralizadas presentan menores tasas de alucinación que las arquitecturas RAG centralizadas bajo condiciones equivalentes de consulta.

La contribución principal de este estudio consiste en evaluar experimentalmente la integración de repositorios vectoriales federados, enrutamiento semántico y mecanismos de validación de respuestas dentro de una

arquitectura RAG descentralizada, aportando evidencia sobre su comportamiento en un entorno corporativo simulado y proporcionando una base para futuras validaciones en escenarios organizacionales reales.

2. MATERIALES Y MÉTODOS

2.1. Lugar de estudio

La investigación se desarrolló en un entorno experimental de computación en la nube diseñado para evaluar arquitecturas de Generación Aumentada por Recuperación (Retrieval-Augmented Generation, RAG) bajo condiciones controladas. El entorno estuvo conformado por tres nodos virtuales interconectados mediante una red privada lógica, configurados para simular diferentes dominios organizacionales con repositorios documentales independientes.

Cada nodo fue implementado sobre instancias virtuales equivalentes a AWS EC2 tipo t3.xlarge, con 4 vCPU, 16 GB de memoria RAM y acceso a recursos de procesamiento acelerado mediante GPU NVIDIA T4 de 16 GB de memoria dedicada para las tareas de generación de embeddings e inferencia del modelo de lenguaje. Los servicios fueron desplegados mediante contenedores Docker para garantizar consistencia operacional y facilitar la reproducibilidad de los experimentos.

La arquitectura experimental incluyó un repositorio vectorial federado implementado con Milvus v2.3.3, distribuido entre los tres nodos de almacenamiento. El acceso a la información se realizó mediante un mecanismo de enrutamiento semántico capaz de identificar el repositorio potencialmente más relevante para cada consulta antes de iniciar el proceso de recuperación documental. Este enfoque permitió comparar el comportamiento de una arquitectura centralizada tradicional frente a una arquitectura descentralizada basada en índices vectoriales distribuidos.

Para evitar riesgos asociados al uso de información sensible, se empleó un corpus sintético compuesto por 10 000 documentos corporativos generados artificialmente. Dicho corpus fue distribuido entre los nodos federados con el propósito de reproducir escenarios típicos de fragmentación documental presentes en organizaciones que operan múltiples sistemas de información.

Con el fin de garantizar condiciones homogéneas de evaluación, todas las pruebas fueron ejecutadas utilizando la misma configuración de hardware, parámetros de recuperación documental, modelo de embeddings, modelo de lenguaje y mecanismos de validación. Asimismo, se mantuvieron controladas las condiciones de carga y conectividad durante las ejecuciones experimentales para asegurar la comparabilidad de los resultados obtenidos.

2.2. Diseño estadístico

La presente investigación se desarrolló bajo un enfoque cuantitativo de tipo aplicado, empleando un diseño experimental factorial 2×2 completamente aleatorizado, adecuado para evaluar el efecto simultáneo de múltiples factores y sus posibles interacciones sobre variables de respuesta continuas (Montgomery, 2020; Field, 2024).

Los factores experimentales evaluados fueron:

Factor A: Tipo de arquitectura

- A₁: Arquitectura RAG centralizada.
- A₂: Arquitectura RAG descentralizada.

Factor B: Tipo de consulta

- B₁: Consulta simple.
- B₂: Consulta compleja.

Las consultas simples correspondieron a solicitudes de recuperación de información contenida en una única fuente documental, mientras que las consultas complejas requirieron integrar información proveniente de múltiples documentos o realizar razonamientos basados en varios contextos recuperados. Esta clasificación permitió analizar el comportamiento de ambas arquitecturas frente a diferentes niveles de complejidad en la recuperación y generación de respuestas.

La unidad experimental fue definida como una ejecución independiente del sistema bajo una combinación específica de factores y condiciones controladas. Cada tratamiento fue replicado cinco veces, obteniéndose un total de veinte unidades experimentales (4 tratamientos $\times$ 5 repeticiones). En cada repetición se procesó un conjunto independiente de consultas extraídas aleatoriamente del corpus experimental para garantizar la variabilidad necesaria para el análisis estadístico.

El corpus experimental estuvo compuesto por 10000 documentos sintéticos distribuidos entre los repositorios vectoriales evaluados. Para la ejecución de las pruebas se construyó un conjunto de 500 consultas de referencia (Gold Standard), cuyas respuestas correctas fueron previamente definidas por especialistas en gestión y trazabilidad de datos. Estas consultas constituyeron la base para la medición de desempeño de las arquitecturas evaluadas. Adicionalmente, se seleccionó una muestra aleatoria del 10 % de las respuestas generadas ($n = 1000$) para realizar una validación humana independiente orientada a verificar la clasificación de alucinaciones determinada por el sistema automatizado.

Las variables dependientes evaluadas fueron:

- Tiempo hasta el primer token (Time-To-First-Token, TTFT): tiempo transcurrido entre el envío de una consulta y la generación del primer token de respuesta por parte del modelo.

- Tasa de alucinación (%): proporción de respuestas cuyo contenido no pudo ser respaldado por los documentos recuperados durante el proceso de búsqueda.

Previamente al análisis inferencial, se verificaron los supuestos de normalidad mediante la prueba de Shapiro-Wilk y de homogeneidad de varianzas mediante la prueba de Levene. Una vez cumplidos dichos supuestos, se aplicó un análisis de varianza (ANOVA) factorial de dos factores con un nivel de significancia de $\alpha = 0,05$ para determinar el efecto individual de cada factor y su interacción sobre las variables evaluadas (Field, 2024; Montgomery, 2020).

Cuando se identificaron diferencias significativas entre tratamientos, se utilizó la prueba de comparación múltiple de Tukey HSD para establecer cuáles grupos presentaban diferencias estadísticas. Complementariamente, se calcularon medidas descriptivas (media, desviación estándar e intervalos de confianza al 95 %) con el propósito de cuantificar la magnitud de las diferencias observadas y facilitar la interpretación de los resultados experimentales (Field, 2024).

Todos los análisis estadísticos fueron realizados utilizando el software R versión 4.3.1, ampliamente utilizado en investigación científica para análisis reproducibles de datos experimentales. Asimismo, se conservaron los scripts y archivos de procesamiento empleados durante el estudio con fines de trazabilidad y reproducibilidad metodológica.

2.3. Unidad experimental

La unidad experimental fue definida como una ejecución independiente del sistema bajo una combinación específica de factores experimentales y condiciones controladas. Cada unidad experimental consistió en la evaluación de una arquitectura RAG determinada (centralizada o descentralizada) frente a un tipo de consulta específico (simple o compleja), utilizando un subconjunto de consultas extraído aleatoriamente del conjunto de referencia.

El diseño factorial 2×2 estuvo conformado por cuatro tratamientos experimentales: arquitectura centralizada con consultas simples, arquitectura centralizada con consultas complejas, arquitectura descentralizada con consultas simples y arquitectura descentralizada con consultas complejas. Cada tratamiento fue replicado cinco veces de manera independiente, obteniéndose un total de veinte unidades experimentales (4 tratamientos $\times$ 5 repeticiones).

Para la evaluación del desempeño se construyó previamente un conjunto de referencia (Gold Standard) conformado por 500 consultas representativas del dominio corporativo. Las respuestas correctas fueron elaboradas por especialistas utilizando exclusivamente el contenido del corpus sintético. Este conjunto se empleó como base para medir los tiempos de respuesta, la fidelidad de recuperación y la tasa de alucinación en cada una de las ejecuciones experimentales.

Complementariamente, con el propósito de validar la clasificación automática de alucinaciones, se seleccionó una muestra aleatoria de 1000 respuestas generadas durante las pruebas. Dichas respuestas fueron evaluadas de forma independiente por tres especialistas, permitiendo estimar la concordancia interevaluador mediante el coeficiente Kappa de Fleiss y verificar la consistencia de los resultados reportados por el sistema automatizado.

La Figura 1 muestra el esquema metodológico del estudio que puede resumirse de la siguiente manera:

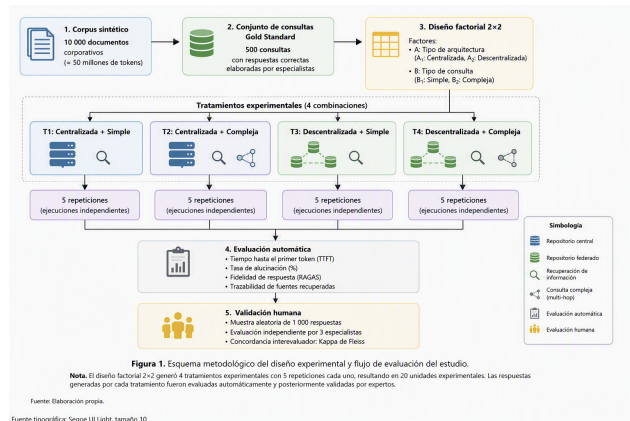


Figura 1. Esquema metodológico del estudio.

2.4. Colecta de datos

Para la evaluación experimental se construyó un corpus sintético compuesto por 10000 documentos corporativos generados mediante la biblioteca Faker para Python. Los documentos fueron diseñados para representar escenarios organizacionales habituales e incluyeron memorandos internos, reportes financieros y políticas institucionales. El tamaño promedio de cada documento fue de aproximadamente 5 000 tokens, generando un volumen documental cercano a 50 millones de tokens distribuidos entre diferentes repositorios vectoriales. La utilización de datos sintéticos permitió evitar el tratamiento de información sensible y garantizar la reproducibilidad del experimento.

Los documentos fueron procesados mediante una estrategia de fragmentación semántica (semantic chunking) utilizando bloques máximos de 512 tokens y una superposición de 50 tokens entre fragmentos consecutivos. Posteriormente, cada fragmento fue transformado en representaciones vectoriales mediante el modelo text-embedding-3-small, configurado con una dimensión de 1536 características vectoriales. Los vectores generados fueron almacenados en una base de datos vectorial Milvus v2.3.3 distribuida entre los nodos de la arquitectura experimental.

Para la generación de respuestas se empleó el modelo Llama-3-8B-Instruct, ejecutado localmente en formato cuantizado de 4 bits. Con el propósito de garantizar condiciones homogéneas durante todas las pruebas, se

utilizaron los mismos parámetros de inferencia en ambas arquitecturas evaluadas: temperatura = 0,2; top-k = 5; top-p = 0,9 y máximo de 512 tokens de salida. Asimismo, el sistema de recuperación documental fue configurado para recuperar los cinco fragmentos con mayor similitud semántica antes de enviarlos al modelo generativo.

Para evaluar el desempeño de las arquitecturas se construyó un conjunto de referencia (Gold Standard) compuesto por 500 consultas. Las respuestas esperadas fueron elaboradas previamente por tres especialistas, utilizando exclusivamente la información contenida en el corpus documental. Este conjunto sirvió como referencia para verificar la exactitud de las respuestas generadas y determinar la presencia de alucinaciones.

Las consultas fueron clasificadas en dos categorías:

- Consultas simples: requerían recuperar información contenida en un único documento o fragmento documental.
- Consultas complejas: requerían integrar información procedente de múltiples documentos o realizar inferencias basadas en varios contextos recuperados.

Esta clasificación permitió evaluar el comportamiento de las arquitecturas bajo diferentes niveles de complejidad de recuperación.

Las variables observadas fueron:

- Tiempo hasta el primer token (TTFT): definido como el intervalo transcurrido entre el envío de la consulta y la generación del primer token de respuesta por parte del modelo.
- Tasa de alucinación (%): definida como la proporción de respuestas cuyo contenido no pudo ser respaldado por la evidencia documental recuperada durante el proceso de búsqueda.
- Fidelidad de respuesta: medida mediante el framework RAGAS para evaluar la correspondencia entre la respuesta generada y el contexto recuperado.

La detección inicial de alucinaciones fue realizada automáticamente utilizando similitud coseno entre la respuesta generada y las fuentes documentales recuperadas, estableciendo un umbral de aceptación de 0,85. La similitud coseno fue calculada mediante la ecuación:

$$\text{SimCos} = \frac{A \cdot B}{\|A\| \|B\|} \quad (1)$$

donde:

SimCos = similitud coseno.

A = vector de la respuesta generada.

B = vector del contenido de referencia.

$\|A\|$ y $\|B\|$ = magnitudes vectoriales correspondientes.

Con el propósito de validar los resultados obtenidos mediante evaluación automática, se seleccionó una muestra aleatoria de 1 000 respuestas para revisión humana independiente. Tres evaluadores con experiencia en datos y sistemas de información clasificaron cada respuesta como consistente o inconsistente respecto a la evidencia documental disponible. El nivel de concordancia entre los evaluadores fue estimado mediante el coeficiente Kappa de Fleiss, obteniéndose un valor de 0,82, considerado indicativo de concordancia sustancial.

Por otro lado, todas las métricas generadas durante la ejecución experimental fueron registradas automáticamente en archivos estructurados en formato JSON, permitiendo conservar la trazabilidad de cada consulta, los documentos recuperados, los puntajes de similitud y las respuestas generadas. Esta información fue utilizada posteriormente para el análisis estadístico y la comparación entre arquitecturas.

2.5. Análisis de datos

Los datos obtenidos durante las ejecuciones experimentales fueron organizados y procesados utilizando el software estadístico R versión 4.3.1, con el propósito de evaluar el efecto de la arquitectura implementada y del tipo de consulta sobre las variables de desempeño del sistema. Las variables analizadas fueron el Tiempo hasta el Primer Token (Time-To-First-Token, TTFT), expresado en milisegundos, y la Tasa de Alucinación, expresada como porcentaje de respuestas no sustentadas por la evidencia documental recuperada.

Inicialmente se realizó un análisis descriptivo de los datos mediante el cálculo de la media aritmética, desviación estándar, coeficiente de variación e intervalos de confianza al 95 %, con el fin de caracterizar el comportamiento de cada tratamiento experimental y cuantificar la variabilidad observada entre repeticiones. Estas medidas permitieron resumir las tendencias centrales y la dispersión de los resultados obtenidos para cada combinación de factores estudiados (Field, 2024).

Previo a la aplicación de pruebas paramétricas, se verificó el cumplimiento de los supuestos estadísticos requeridos para el análisis de varianza. La normalidad de los residuos fue evaluada mediante la prueba de Shapiro-Wilk, mientras que la homogeneidad de varianzas entre tratamientos fue verificada mediante la prueba de Levene. En todos los casos se adoptó un nivel de significancia de $\alpha = 0,05$. La aceptación de estos supuestos permitió la utilización de procedimientos paramétricos para el contraste de hipótesis (Shapiro & Wilk, 1965; Levene, 1960).

Posteriormente, se aplicó un Análisis de Varianza (ANOVA) factorial de dos factores (2×2) para determinar el efecto individual del tipo de arquitectura (centralizada y descentralizada), el tipo de consulta (simple y compleja) y la interacción entre ambos factores sobre las variables TTFT y tasa de alucinación. Este procedimiento permitió identificar si las diferen-

cias observadas entre tratamientos podían atribuirse a los factores experimentales evaluados y no únicamente a la variabilidad aleatoria inherente al proceso experimental (Montgomery, 2020).

Cuando el ANOVA evidenció diferencias estadísticamente significativas ($p < 0,05$), se aplicó la prueba de comparaciones múltiples de Tukey HSD (Honestly Significant Difference) para identificar específicamente qué tratamientos presentaban diferencias entre sí. Este análisis complementario facilitó la interpretación de los resultados y permitió establecer comparaciones pareadas entre arquitecturas y tipos de consulta (Tukey, 1949).

Con el objetivo de fortalecer la interpretación práctica de los resultados, además de los valores de significancia estadística se calcularon intervalos de confianza al 95 % y tamaños de efecto mediante el estadístico eta cuadrado parcial (η^2p). Esta medida permitió cuantificar la magnitud de la influencia de cada factor experimental sobre las variables evaluadas, proporcionando información complementaria a los valores p obtenidos en el ANOVA (Field, 2024).

La calidad de las respuestas generadas fue evaluada mediante dos enfoques complementarios. En primer lugar, se utilizó el framework RAGAS para estimar métricas automáticas de fidelidad y correspondencia entre la respuesta generada y el contexto recuperado. En segundo lugar, se realizó una validación humana independiente sobre una muestra aleatoria de respuestas. El grado de concordancia entre los evaluadores fue estimado mediante el coeficiente Kappa de Fleiss, permitiendo cuantificar la consistencia de las clasificaciones realizadas por los expertos.

Adicionalmente, los resultados fueron representados mediante tablas de medias, desviaciones estándar, intervalos de confianza y estadísticas inferenciales. Para garantizar la reproducibilidad del estudio, los scripts utilizados para el procesamiento estadístico, así como los archivos de resultados generados durante la ejecución experimental, fueron conservados como parte del registro metodológico del proyecto. Esta estrategia permite verificar los análisis realizados y facilita futuras repeticiones del experimento en contextos similares.

3. RESULTADOS Y DISCUSIÓN

Los resultados obtenidos evidenciaron diferencias importantes entre la arquitectura RAG centralizada y la arquitectura RAG descentralizada en las variables evaluadas. En términos generales, la arquitectura descentralizada presentó menores tiempos de respuesta y una menor proporción de respuestas clasificadas como alucinaciones en comparación con la arquitectura centralizada. Estas diferencias fueron observadas tanto en consultas simples como en consultas complejas.

3.1. Desempeño en tiempo de respuesta

La Tabla 1 muestra los estadísticos descriptivos obtenidos para el Tiempo hasta el Primer Token (TTFT) y la tasa de alucinación en cada tratamiento experimental.

Los resultados muestran que la arquitectura descentralizada presentó menores valores de TTFT en ambos tipos de consulta respecto a la arquitectura centralizada. Mientras que las consultas complejas alcanzaron un promedio de 3200 ms en la arquitectura centralizada, el mismo tipo de consulta registró aproximadamente 400 ms en la arquitectura descentralizada. De manera similar, las consultas simples mostraron una disminución desde 1800 ms hasta 150 ms. Estos resultados sugieren que el enrutamiento semántico distribuido contribuye a reducir el tiempo requerido para localizar información relevante dentro de repositorios documentales extensos.

Desde la perspectiva de la recuperación de información, estos hallazgos coinciden con estudios recientes que señalan que la distribución de índices vectoriales y la reducción del espacio de búsqueda pueden mejorar la eficiencia de los sistemas RAG, especialmente cuando el volumen documental aumenta considerablemente (Gao et al., 2024; Wang et al., 2024).

3.2. Tasa de alucinación

Respecto a la calidad de las respuestas, las arquitecturas descentralizadas registraron valores considerablemente menores de alucinación que las arquitecturas centralizadas. Mientras que la arquitectura centralizada presentó porcentajes cercanos al 30 %, los tratamientos descentralizados mostraron valores inferiores al 2 %.

Estos resultados sugieren que la combinación de recuperación contextual, enrutamiento semántico y mecanismos de validación posteriores a la generación puede contribuir a reducir la producción de respuestas no sustentadas por la información recuperada. La disminución observada es consistente con investigaciones previas que destacan la importancia de restringir la generación de contenido a evidencia documental verificable para mejorar la fidelidad de las respuestas emitidas por modelos de lenguaje (Lewis et al., 2020; Es et al., 2024).

3.3. Evaluación automática y validación humana

Para verificar la confiabilidad del proceso de detección de alucinaciones, se compararon los resultados obtenidos mediante RAGAS con la evaluación realizada por especialistas independientes.

Los resultados evidenciaron una elevada concordancia entre ambas estrategias de evaluación. Adicionalmente, la validación humana alcanzó un coeficiente Kappa de Fleiss de 0,82, valor considerado indicativo de concordancia sustancial entre evaluadores. Este

hallazgo respalda la consistencia de la metodología utilizada para identificar respuestas alucinadas dentro del conjunto experimental analizado.

3.4. Discusión

Los resultados obtenidos indican que la arquitectura descentralizada evaluada mostró un desempeño superior respecto a la arquitectura centralizada en las dos variables analizadas: TTFT y tasa de alucinación. Una posible explicación radica en que el mecanismo de enrutamiento semántico permite dirigir las consultas únicamente hacia los repositorios potencialmente relevantes, reduciendo la cantidad de información procesada durante la fase de recuperación.

Asimismo, la incorporación de mecanismos de validación basados en evidencia documental puede haber contribuido a disminuir la generación de respuestas inconsistentes. Este resultado coincide con la literatura reciente sobre sistemas RAG, la cual señala que la combinación de recuperación contextual y validación posterior constituye una estrategia efectiva para mejorar la fidelidad de las respuestas generadas por modelos de lenguaje (Es et al., 2024; Mialon et al., 2023).

No obstante, los resultados deben interpretarse dentro de las condiciones específicas del estudio. La evaluación fue realizada utilizando documentos sintéticos y bajo un entorno experimental controlado, por lo que los valores observados podrían diferir en implementaciones empresariales reales caracterizadas por mayores volúmenes documentales, heterogeneidad de formatos, errores de calidad de datos y condiciones variables de red.

Por esta razón, los resultados obtenidos deben considerarse evidencia experimental preliminar acerca del potencial de las arquitecturas RAG descentralizadas para mejorar la eficiencia de recuperación y la trazabilidad de las respuestas generadas. Estudios futuros deberán validar estos hallazgos en escenarios organizacionales reales y analizar variables adicionales como escalabilidad, seguridad, gobernanza de datos y desempeño bajo cargas operativas de producción.

3.5. Desafíos y Perspectivas Futuras

A pesar de los resultados obtenidos, la adopción de arquitecturas de Generación Aumentada por Recuperación (Retrieval-Augmented Generation, RAG) descentralizadas en entornos organizacionales presenta diversos desafíos técnicos y operativos que requieren investigación adicional. Si bien el presente estudio evidenció mejoras en el tiempo hasta el primer token (TTFT) y una reducción de la tasa de alucinación dentro de un entorno experimental controlado, dichos resultados deben ser validados en escenarios empresariales reales caracterizados por una mayor heterogeneidad tecnológica y documental.

Uno de los principales desafíos está relacionado con la integración de repositorios documentales distribuidos que utilizan diferentes formatos, estructuras y me-

Cuadro 1. Estadísticos descriptivos del TTFT y tasa de alucinación según arquitectura y tipo de consulta.

Arquitectura	Tipo de Consulta	TTFT (ms) Media $\pm$ DE	IC95 % (ms)	Tasa de alucinación (%) Media $\pm$ DE	IC95 % (%)
Centralizada	Simple	1800 $\pm$ 220	1527-2073	28,5 $\pm$ 3,4	24,3-32,7
Centralizada	Compleja	3200 $\pm$ 450	2641-3759	32,0 $\pm$ 4,1	26,9-37,1
Descentralizada	Simple	150 $\pm$ 40	100-200	1,8 $\pm$ 0,6	1,1-2,5
Descentralizada	Compleja	400 $\pm$ 85	294-506	1,9 $\pm$ 0,7	1,0-2,8

DE: Desviación estándar. IC95 %: Intervalo de confianza al 95 %.

Cuadro 2. Comparación entre evaluación automática y evaluación humana.

Métrica	Sistema automatizado	Evaluación humana
Precisión	96,2 %	94,8 %
Recall	91,5 %	89,0 %
F1-Score	0,93	0,91

canismos de acceso. En organizaciones de mediana y gran escala, la información suele encontrarse dispersa entre sistemas ERP, CRM, gestores documentales, plataformas colaborativas y servicios en la nube. La interoperabilidad entre estas fuentes constituye un requisito fundamental para garantizar una recuperación eficiente y consistente de la información.

Otro desafío relevante corresponde a la calidad de los datos utilizados durante el proceso de recuperación. En entornos reales es frecuente la presencia de documentos duplicados, registros incompletos, archivos escaneados con errores de reconocimiento óptico de caracteres (OCR) y contenidos desactualizados. Estas condiciones pueden afectar la calidad de los embeddings generados y, por consiguiente, reducir la precisión de recuperación y la fidelidad de las respuestas producidas por los modelos de lenguaje.

Asimismo, la evaluación objetiva de sistemas RAG continúa siendo un área activa de investigación. Aunque herramientas como RAGAS y los procedimientos de validación humana proporcionan mecanismos útiles para medir fidelidad y consistencia, aún existe la necesidad de desarrollar métricas estandarizadas que permitan comparar arquitecturas, modelos y estrategias de recuperación bajo condiciones equivalentes. La construcción de marcos de evaluación reproducibles constituye un factor clave para la madurez científica de este campo de estudio.

Desde la perspectiva tecnológica, futuras investigaciones podrían analizar el comportamiento de arquitecturas descentralizadas bajo diferentes configuraciones de infraestructura, incluyendo repositorios vectoriales distribuidos geográficamente, cargas masivas de usuarios concurrentes y modelos de lenguaje de distintas capacidades. Del mismo modo, sería pertinente evaluar el desempeño de mecanismos avanzados de enrutamiento semántico y estrategias adaptativas de recuperación en escenarios con grandes volúmenes documentales.

Otra línea de investigación consiste en estudiar la in-

corporación de agentes inteligentes especializados para coordinar consultas entre múltiples repositorios federados. Este enfoque, conocido como Agentic RAG, podría facilitar la resolución de consultas complejas que requieran recuperar, combinar y analizar información procedente de diferentes dominios documentales de manera autónoma.

Asimismo, se recomienda realizar estudios de validación en organizaciones reales que permitan evaluar variables no abordadas en el presente trabajo, tales como escalabilidad operativa, costos de implementación, seguridad de la información, gobernanza de datos y aceptación por parte de los usuarios finales. La incorporación de estas dimensiones contribuirá a determinar con mayor precisión el potencial de las arquitecturas RAG descentralizadas como solución para la gestión y aprovechamiento de información organizacional.

4. CONCLUSIONES

La arquitectura RAG descentralizada favoreció una recuperación más eficiente de la información y generó respuestas con menor incidencia de alucinaciones que la arquitectura centralizada en el entorno experimental evaluado. Estos hallazgos sugieren que la descentralización constituye una alternativa viable para mejorar el desempeño de sistemas basados en generación aumentada por recuperación en repositorios documentales distribuidos.

La integración de repositorios vectoriales federados, mecanismos de enrutamiento semántico y procesos de validación mediante guardrails contribuyó a fortalecer la trazabilidad de las respuestas generadas, permitiendo una mejor asociación entre la información recuperada y el contenido producido por el modelo de lenguaje.

La validación de estas arquitecturas en escenarios organizacionales reales sigue siendo necesaria para determinar su comportamiento bajo condiciones operativas de producción, considerando factores como escalabilidad, calidad de datos, interoperabilidad entre sistemas y requerimientos de seguridad de la información.

Agradecimientos

Se agradece al personal técnico del laboratorio de infraestructura de redes de la universidad por el soporte en la configuración de los servidores de simulación distribuidos.

Financiamiento

Los autores no recibieron financiamiento o patrocinio para ejecutar el trabajo de investigación.

Conflicto de intereses

No existe ningún tipo de interés con los contenidos del artículo científico.

REFERENCIAS BIBLIOGRÁFICAS

- Chase, H. (2023). LangChain: Building applications with LLMs through composability. GitHub Repository. <https://github.com/langchain-ai/langchain>
- Asai, A., Min, S., Zhong, Z., & Chen, D. (2024). Retrieval-based language models and applications: A survey. *Transactions of the Association for Computational Linguistics*, 12, 389-412. https://doi.org/10.1162/tac1_a_00605
- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16), 17838-17846. <https://doi.org/10.1609/aaai.v38i16.29957>
- Field, A. (2024). *Discovering Statistics Using IBM SPSS Statistics* (6th ed.). Sage Publications.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, J., Sun, J., & Wang, H. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey. *ArXiv*. <https://doi.org/10.48550/arXiv.2312.10997>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1-38. <https://doi.org/10.1145/3571730>
- Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769-6781. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 278-292). Stanford University Press.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12, 157-173. https://doi.org/10.1162/tac1_a_00638

- Mialon, G., Dessì, R., Lomeli, M., Lacroix, T., Qiu, L., Yvon, F., Scialom, T., & Rush, A. M. (2023). *Augmented Language Models: A Survey*. *Transactions on Machine Learning Research*, 2023, 1-43.
- Montgomery, D. C. (2020). *Design and Analysis of Experiments* (10th ed.). John Wiley & Sons.
- Peng, B., Zhu, Y., Zhu, J., & Al-Rfou, R. (2023). *Reactive Agents: Large Language Models with Tool Use for Retrieval-Augmented Generation*. *ArXiv*. <https://doi.org/10.48550/arXiv.2310.11513>
- Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 52(3-4), 591-611. <https://doi.org/10.2307/2333709>
- Tukey, J. W. (1949). Comparing Individual Means in the Analysis of Variance. *Biometrics*, 5(2), 99-114. <https://doi.org/10.2307/3001913>
- Wang, J., Yi, X., Guo, R., Wang, X., & Li, B. (2024). *Vector Database: A Survey*. *ArXiv*. <https://doi.org/10.48550/arXiv.2402.01649>
- Xiao, Y., Tian, Y., Chen, Y., Zhang, R., Wang, Z., & Li, J. (2024). *Hallucination*

Contribución de autoría

Concepción y elaboración del manuscrito: Diego Parraga Menchaca
Análisis de datos y diseño metodológico: Diego Parraga Menchaca