

ARTÍCULO DE INVESTIGACIÓN

Aplicación web con arquitectura RAG y LLaMa 3.2 para consultas de soporte técnico en CNEL

Web application with RAG system and LLaMa 3.2 for technical support queries at CNEL

DOI: <https://doi.org/10.61154/metanoia.v12i1.4282>

Wilson Josue González Ortega¹

Mariuxi Paola Zea Ordoñez²

Johnny Paul Novillo Vicuña³

Alex Roberto Ortiz Ortega⁴

¹**E-mail:** wgonzalez6@utmachala.edu.ec **Afiliación:** Universidad Técnica de Machala. El Oro, Machala, Ecuador. **ORCID:** <https://orcid.org/0000-0003-4701-981X>

²**E-mail:** mzea@utmachala.edu.ec **Afiliación:** Universidad Técnica de Machala. El Oro, Machala, Ecuador. **ORCID:** <https://orcid.org/0000-0001-8860-6282>

³**E-mail:** jnovillo@utmachala.edu.ec **Afiliación:** Universidad Técnica de Machala. El Oro, Machala, Ecuador. **ORCID:** <https://orcid.org/0000-0002-4915-3441>

⁴**E-mail:** alex.ortiz@cnel.gob.ec **Afiliación:** Empresa eléctrica pública estratégica Corporación Nacional de Electricidad. El Oro, Machala, Ecuador. **ORCID:** <https://orcid.org/0009-0009-6982-8440>

Recibido: 09/10/2025

Revisado: 19/10/2025

Aprobado: 09/11/2025

Publicado: 01/01/2026

RESUMEN

La gestión del conocimiento técnico en la Corporación Nacional de Electricidad (CNEL EP) del Ecuador se ve obstaculizada por la dispersión de la información en múltiples documentos, lo que ralentiza la resolución de problemas en el área de soporte técnico. Para enfrentar este desafío, se desarrolló una aplicación web que implementa una arquitectura de Generación Aumentada por Recuperación (RAG) junto con el modelo de lenguaje Llama 3.2. La metodología de desarrollo siguió el proceso CRISP-DM, asegurando un enfoque estructurado desde la comprensión del negocio hasta el despliegue. El sistema permite cargar documentos institucionales (PDF, Word, Excel), los cuales son procesados, convertidos en vectores semánticos mediante el modelo nomic-embed-text y almacenados en una base de datos vectorial FAISS para una recuperación eficiente, al realizar una consulta, el sistema recupera los fragmentos de texto más relevantes y los utiliza como contexto para que el modelo Llama 3.2, ejecutado localmente con Ollama, genere una respuesta precisa y contextualizada. Las medidas de la evaluación por la norma ISO/IEC 9126 indicaron un alto nivel de desempeño en funcionalidad (4.5/5) y usabilidad (4.8/5), confirmando el potencial del modelo para la centralización del conocimiento y para la producción de respuestas fiables. Del mismo modo que el despliegue local garantiza la seguridad de los datos, la eficiencia fue vista como un área de mejora, dado que los tiempos de respuesta promedios fueron de 8.38 segundos. La herramienta se consolida como un apoyo inteligente para agilizar el acceso a la información técnica institucional.

Descriptor: Tecnología de la información, gestión de la información, protección de datos, inteligencia artificial.

ABSTRACT

Technical knowledge management at Ecuador's National Electricity Corporation (CNEL EP) is hampered by the dispersion of information across multiple documents, which slows problem resolution in the technical support area. To address this challenge, a web application was developed that implements a Recovery Augmented Generation (RAG) architecture along with the Llama 3.2 language model. The development methodology followed the CRISP-DM process, ensuring a structured approach from business understanding to deployment. The system allows the upload of institutional documents (PDF, Word, Excel), which are processed,

converted into semantic vectors using the nomic-embed-text model, and stored in a FAISS vector database for efficient retrieval. When a query is performed, the system retrieves the most relevant text fragments and uses them as context for the Llama 3.2 model, run locally with Ollama, to generate an accurate and contextualized response. The ISO/IEC 9126 evaluation metrics indicated a high level of performance in functionality (4.5/5) and usability (4.8/5), confirming the model's potential for centralizing knowledge and producing reliable responses. While local deployment ensures data security, efficiency was seen as an area for improvement, with average response times of 8.38 seconds. The tool is established as an intelligent support for streamlining access to institutional technical information.

Descriptors: Information technology, information management, data protection, artificial intelligence.

INTRODUCCIÓN

En Ecuador se maneja la empresa eléctrica pública estratégica Corporación Nacional de Electricidad (CNEL EP), encargada del servicio público de distribución y comercialización de energía eléctrica en gran parte del territorio nacional del Ecuador. CNEL EP tiene la responsabilidad de satisfacer la demanda energética conforme a las disposiciones del marco normativo del sector eléctrico nacional (CNEL EP, 2016).

La dispersión de la información en el área de soporte técnico en la compañía de CNEL EP dificulta el acceso rápido del conocimiento institucional valioso. El personal de soporte debe consultar manuales, normativas y registros almacenados en diversos documentos digitales o estar enviando correos para ayuda en consulta a los del área de arquitectura y aplicaciones, lo que retrasa la solución de problemas y sobrecarga los procesos operativos. Según Alghanemi & Al Mubarak (2022), la gestión del conocimiento (KM) es el proceso de capturar, distribuir y utilizar eficientemente el conocimiento organizacional, y la inteligencia artificial se ha convertido en un elemento clave en los procesos de generación de conocimiento en las empresas. Estudios similares han demostrado que la implementación de chatbots de inteligencia artificial (IA) puede mejorar significativamente en el soporte técnico, por ejemplo, (Maruri et al., 2025) reportan una reducción del 30% en tiempos de atención y un aumento del 25% en la satisfacción del cliente tras integrar un chatbot de IA en el CRM de soporte. (Alghanemi & Al Mubarak, 2022, p. 367) destacan que “la inteligencia artificial es una herramienta poderosa que apoya el conocimiento en todos los aspectos, y esta colaboración respalda el proceso KM

con todos los avances e innovaciones en la tecnología integrada por la IA”. Estos resultados sugieren que una herramienta conversacional inteligente que centralice el conocimiento institucional podría optimizar la búsqueda de información en el área de soporte técnico en CNEL EP.

La arquitectura de Generación Aumentada por Recuperación (RAG) resulta adecuada para enfrentar este reto, ya que consiste en combinar un modelo de lenguaje con un módulo de recuperación de información que consulta directamente la base de conocimiento institucional antes de generar la respuesta. En otras palabras, extiende las capacidades de un modelo de lenguaje de gran escala (LLM) a los dominios específicos de la empresa sin necesidad de reentrenar el modelo (Sánchez García, 2024). De modo, el modelo recupera información autorizada de documentos internos y la utiliza para generar respuestas de manera más precisa y actualizadas, como señalan fuentes especializadas, la arquitectura RAG permite incorporar “contenido empresarial propietario” en la formulación de las respuestas (HeidiSteen, 2025). En consecuencia, los chatbots basados en RAG pueden ofrecer respuestas más precisas y contextualizadas al acceder a base de conocimiento.

Para el componente generativo se opta por Llama 3.2 (3B), un modelo de lenguaje abierto optimizado para diálogo y recuperación de información, combina un buen rendimiento con la posibilidad de los datos de CNEL EP. Herramientas como Ollama, simplifican la forma de implementar Llama 3.2 al ofrecer entornos preconfigurados. De esta manera se puede obtener respuestas rápidas y de calidad sin requerir de una infraestructura compleja. Si integramos Llama 3.2 dentro el flujo RAG, el chatbot que se genera puede llegar a obtener respuestas altamente contextuales, esto es gracias a los fragmentos relevantes de los documentos institucionales.

Este trabajo presenta el desarrollo de un sistema basado en una aplicación web con arquitectura RAG y Llama 3.2 para el área de soporte técnico de CNEL EP. La pregunta de investigación es: ¿Cómo puede un sistema basado en RAG y Llama 3.2 unificar la información técnica que está dispersa y contestar correctamente a las consultas del soporte técnico, generando respuestas pertinentes de acuerdo con el contexto? Nuestra hipótesis es que un sistema basado en RAG y Llama 3.2 permitirá centralizar el conocimiento técnico de la institución en un repositorio semántico eficiente, dar acceso a la información de forma más rápida e incrementar la velocidad en la atención al usuario. El desarrollo del proyecto seguirá la metodología CRISP-DM, mediante la que se estructuran las fases del proyecto, desde la comprensión del problema hasta

la fase de despliegue de la solución.

El aspecto fundamental es el desarrollo de una aplicación web que va ser un chatbot inteligente que permite el acceso rápido y fiable a la información técnica de los documentos del área de soporte técnico. Como alcance, la aplicación web permitirá: (I) cargar los documentos en múltiples formatos de archivo (PDF, Word, Excel), (II) procesar la información mediante técnicas de extracción, limpieza y segmentación, (III) integrar los documentos a un índice vectorial para búsqueda semántica, y (IV) dar respuestas contextualizadas a partir de un modelo LLM altamente eficaz.

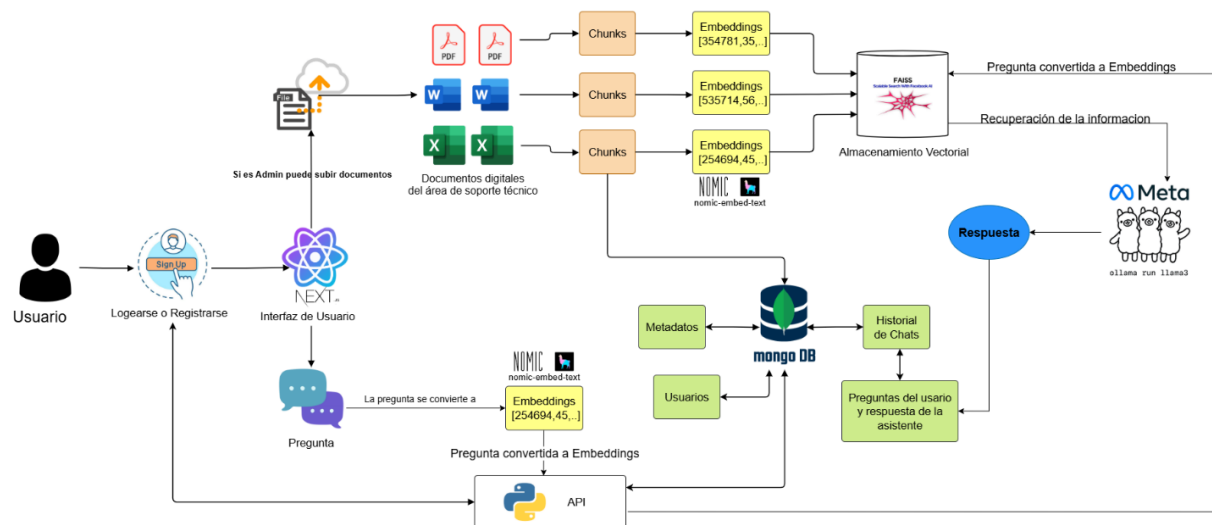
En este mismo sentido, la investigación, además de tratar de optimizar la eficacia en la accesibilidad del conocimiento técnico dentro de CNEL EP, pretende creó un marco de referencia para establecer pautas metodológicas que puedan ser replicadas en otras instituciones que enfrenten situaciones similares, aportando de esta forma avances en la gestión de conocimiento organizacional fundamentado en IA.

METODOLOGÍA

La presente investigación se encuentra centrada en la creación de una aplicación web apoyada en una arquitectura RAG, orientada a la mejora de los procesos que tienen lugar en el área de soporte técnico de CNEL EP mediante la realización de consultas contextualizadas a documentos digitales institucionales en el contexto de la utilización del modelo de lenguaje Llama 3.2 para la generación de las respuestas, utilizando para ello la metodología CRISP-DM (Cross Industry Standard Process for Data Mining). Esta metodología para el descubrimiento de conocimiento en bases de datos, está organizada como un proceso jerárquico en seis fases básicas: comprensión del negocio, comprensión de datos, preparación de los datos, modelado, evaluación y despliegue (Ahumada et al., 2020). Para el caso de CNEL EP, la aplicación de este modelo resulta adecuado, ya que facilita la organización del trabajo desde la identificación del problema hasta el despliegue del sistema, garantizando la transformación e interpretación adecuada de los documentos técnicos no estructurados gestionados institucionalmente.

Arquitectura General del Sistema

La arquitectura del sistema, representada en la Figura 1, integra múltiples componentes tecnológicos que operan de forma coordinada para brindar una experiencia de usuario y segura.

Figura 1*Arquitectura general de la aplicación web basado en RAG***Fuente:** Los autores

El flujo de operación se inicia con el módulo de autenticación, donde los usuarios deben registrarse o iniciar sesión con sus credenciales. Una vez autenticado, aquellos usuarios con perfil de administrador poseen permisos adicionales para cargar los documentos digitales en múltiples formatos (PDF, Word, Excel) para su procesamiento, donde el contenido es extraído y indexado en la base de datos vectorial FAISS, una librería que búsqueda de similitud entre vectores, ofreciendo métodos de indexación, comprensión y clustering que permite gestionar grandes volúmenes de embeddings y realizar búsquedas aproximadas de vecinos más cercanos (Douze et al., 2025), tanto los administradores como los usuarios regulares puede realizar consultas a través de la interfaz web, la cual es transformada en un vector de embeddings para realizar una búsqueda semántica en el almacenamiento vectorial FAISS. Los fragmentos de texto relevantes son recuperados y, junto con el historial de conversación almacenado en MongoDB, se envían al modelo LLaMa 3.2 ejecutando localmente en Ollama, para generar una respuesta contextualizada. Finalmente, cada consulta y su respuesta generada se almacenan en el historial de conversación para mantener el contexto en interacciones futuras. Este flujo adopta una arquitectura RAG, el cual combina mecanismos de recuperación de información y generación de texto para mejorar la precisión de las respuestas. En este enfoque, los documentos se fragmentan en chunks y se transforman en representaciones vectoriales para realizar búsquedas semánticas, de modo que los fragmentos más relevantes se incorporan al prompt del

modelo antes de generar la respuesta final (Sánchez García, 2024).

Comprensión del Negocio

El proceso se inicia con la fase de comprensión del negocio, cuya finalidad es comprender los objetivos y requerimientos desde una perspectiva organizacional, para posteriormente traducir este conocimiento en una definición concreta del problema técnico y un plan preliminar destinado a alcanzar los objetivos establecidos (Calatrava Hurtado, 2019).

Se identificó que el principal problema del que adolece CNEL EP es la dispersión que tiene la información del área de soporte técnico en documentos no estructurados, lo que conlleva complicaciones de acceso a la información y, de ahí, dificultades en la operatividad y toma de decisiones, estableciéndose por lo tanto como objetivo principal el desarrollo de una solución basada en inteligencia artificial para centralizar, recuperar y sintetizar información útil de documentos digitales a través de una interfaz web accesible, usando para ello la metodología CRISP-DM por su capacidad de organizar proyectos desde la comprensión del problema hasta la implementación. Añadiendo además reuniones con expertos del CNEL EP para acordar criterios de éxito y definir el alcance del proyecto siguiendo las necesidades organizacionales, considerando que la implementación de herramientas tipo IA como chatbot con procesamientos de lenguaje natural (PLN) han demostrado ser efectivas para aumentar la accesibilidad de la información y optimizar procesos operativos (Skinner Barbosa & García Díaz, 2025).

Se identificaron los siguientes parámetros que requería la empresa:

- Centralizar el conocimiento técnico disperso en múltiples formatos documentales.
- Proporcionar respuestas contextuales al personal técnico.
- Implementar un sistema de autenticación para controlar el acceso a la información y tanto alguien que pueda administrar los documentos con los que se va ser entrenado el sistema.
- Garantizar la privacidad de los datos mediante la utilización de modelos locales.
- Mantener un historial de conversación para el seguimiento y la mejora continua.

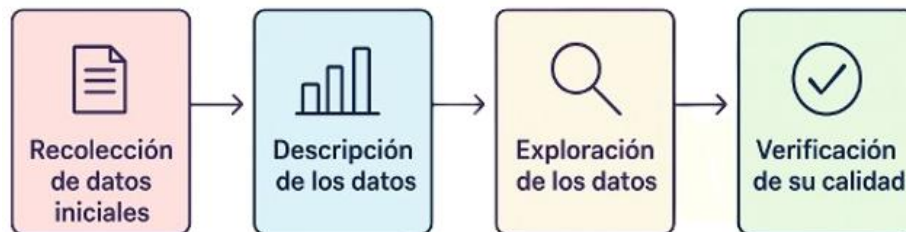
Etapas de comprensión de los datos

En la fase de comprensión de los datos, se lleva a cabo la recolección inicial de la información a propósito de establecer un primer acercamiento al problema, lo que permite familiarizar con ellos, identificar su calidad y determinar relaciones evidentes para formular las primeras hipótesis. Entre las principales tareas de esa fase se incluye: recolección de datos iniciales, la

descripción de los datos, la exploración de esos y a verificación de su calidad (Tamayo Cárdenas & Villavicencio Martínez, 2018).

Figura 2

Fase de comprensión de los datos



Fuente: Los autores

Para el desarrollo de la aplicación web de CNEL EP, se inició con la recolección de fuentes documentales no estructuradas esenciales para el dominio técnico de CNEL EP, que incluyeron documentos en formatos PDF (manuales de procedimientos, normativas y guías), archivos Word (informes y especificaciones) y hojas de cálculo Excel (inventarios y registros de mantenimiento). El análisis de estos materiales reveló características significativas: variedad en formatos, contenido principalmente no estructurado, presencia de tanto texto digitalizado como imágenes que requieren procesamiento OCR.

La exploración detallada permitió identificar que la información está dispersa en diferentes repositorios, también el uso de términos técnicos muy especializados en CNEL EP. La verificación de calidad, se encontró que los documentos habían sido redactados de diferentes maneras y en distintos momentos, por lo que el contenido técnico había sido previamente validado por expertos, era importante considerar que alguna información podría estar desactualizada al momento de implementar el sistema.

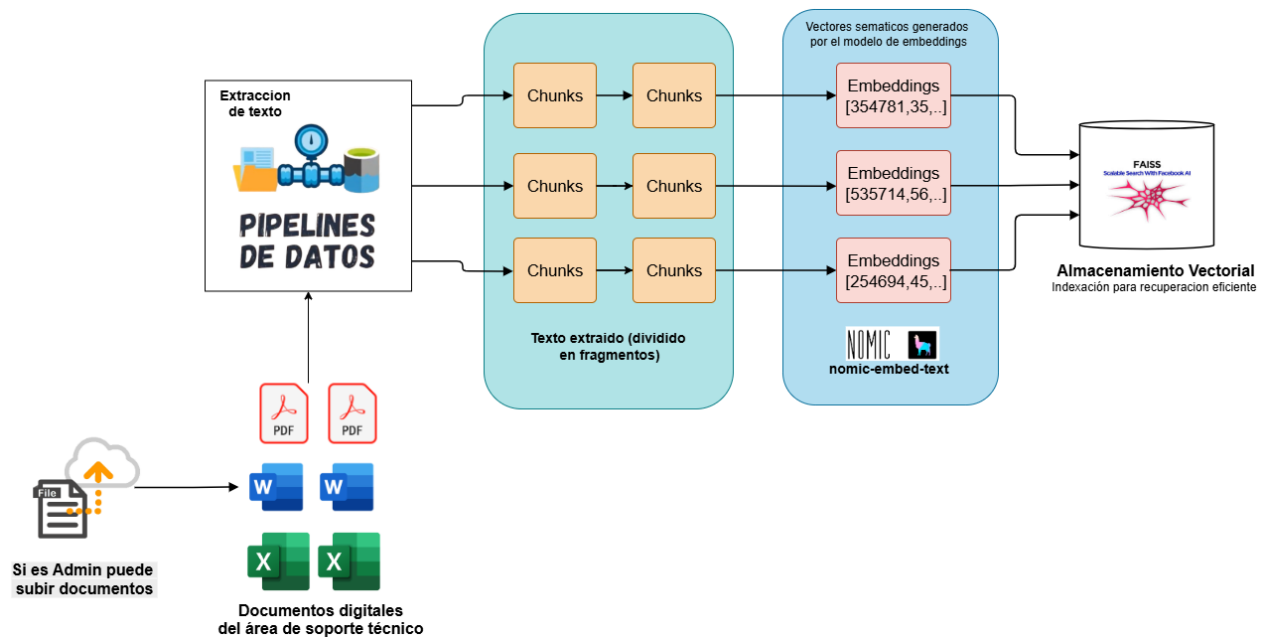
Se realizó un análisis exploratorio que reveló la necesidad de implementar un pipeline robusto de extracción y normalización de texto en la aplicación de CNEL EP se debe a la naturaleza dispersa y no estructurada de la información recopilada (PDF, Word, Excel, con texto e imágenes). Investigaciones han demostrado que, en este tipo de escenarios, un pipeline permite automatizar etapas críticas como la lectura de documentos, el procesamiento OCR y la normalización del texto, garantizando datos consistentes y utilizables para las fases posteriores de análisis (Bregón Bregón & Silvestre Vilches, 2021).

Preparación de los Datos

La fase de preparación de los datos es fundamental dentro del ciclo CRISP-DM, ya que transforma los datos brutos iniciales en un conjunto de alta calidad y adecuado para el modelado, se prepararon los datos para el modelado. Este es el procedimiento más extenso ya que se debe analizar qué datos son relevantes para el modelo para luego realizar la limpieza de datos la cual es una tarea muy larga (Tamariz Milla et al., 2025).

La preparación de los datos se realizó mediante un pipeline estructurado que transforma los documentos en un formato utilizable por el sistema de inteligencia artificial. Como se muestra en la Figura 3, el proceso inicia cuando el administrador carga los documentos digitales al sistema, estos documentos pasan una fase de extracción de texto donde se convierte su contenido textual y se divide en fragmentos manejables.

Posteriormente, estos fragmentos de texto son transformados en presentaciones numéricas mediante el modelo de embeddings nomic-embed-text que es un modelo de embeddings de texto que Nomic desarrolló como base para entrenar otros modelos y que destaca por su rendimiento de generación de representaciones numéricas para texto (Nussbaum et al., 2024), que se lo puede descargar de manera local mediante Ollama y con este modelo se genera vectores semánticos que capturan el significado del contenido técnico. Estos embeddings se almacenan en una base de datos vectorial que permite la recuperación eficiente de información durante la operación del sistema, mientras que los metadatos asociados a cada fragmento incluyendo un identificador único, tipo de documento, área técnica y fecha de actualización se almacenaran en MongoDB para una gestión de archivos.

Figura 3*Proceso de preparación de datos***Fuente:** Los Autores

El conjunto de procesos del pipeline que se implementa, ha integrado procesos de limpieza y normalización de texto; control de los caracteres especiales del español; segmentación del texto, actuando de forma cuidadosa a la hora de mantener el contexto técnico de la información. Cada uno de estos fragmentos es etiquetado con metadatos que permiten la posibilidad de recuperar el documento original, garantizando que las respuestas generadas por el sistema puedan referenciar las fuentes institucionales que corresponden, a saber, la suya. En este sentido, la sistemática de preparación de los datos, constituye la base fundamental que permitirá al sistema RAG, acceder y poder utilizar de una forma adecuada el conocimiento técnico institucional para generar respuestas precisas y en su contexto.

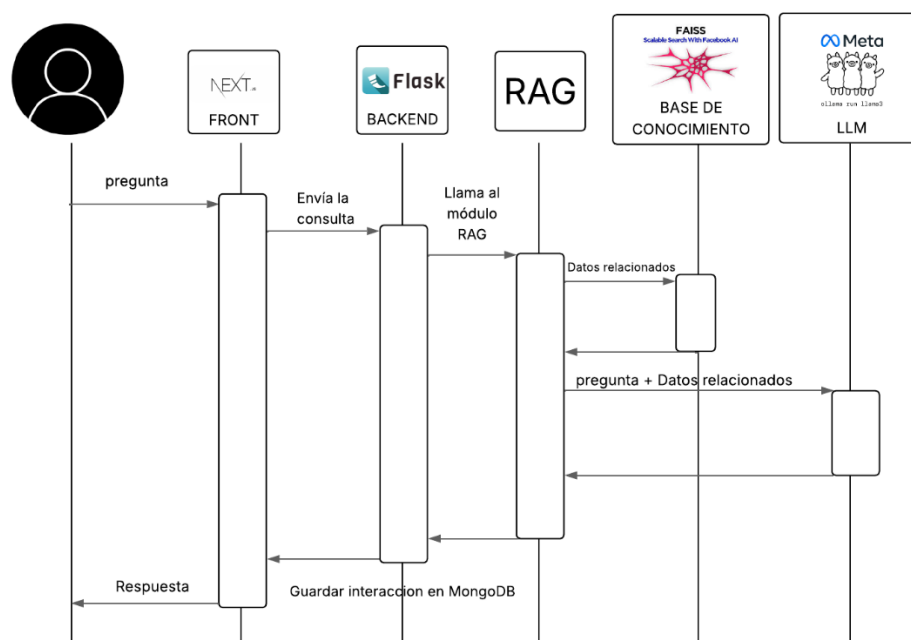
Modelado

La etapa de modelado se desarrolla en el corazón del proceso CRISP-DM, aplicándose técnicas algorítmicas específicas a los datos preparados; se desarrollan modelos que darán respuesta a los objetivos del negocio. En esta etapa se escoge la técnica de modelo real que se usará. Puede ser que se haya preseleccionado una herramienta en la fase de comprensión del negocio, pero es aquí donde se determinará qué técnica de modelado a utilizar y cómo se fijarán los parámetros de la misma (Calatrava Hurtado, 2019).

La arquitectura de tipo RAG se ha considerado para la realización del sistema propuesto, tal como se representa en la Figura 4, que servirá para combinar la tarea de recuperar información con la tarea de obtener la respuesta en lenguaje natural de manera eficiente. Esta arquitectura combina componentes especializados en embeddings, gestión de contexto y generación de respuestas, soportados por un backend robusto y una interfaz moderna. El componente central es un modelo de lenguaje Llama 3.2 (específicamente la versión 3B) fue evaluado para generación de preguntas automáticas, demostrando mejor rendimiento en tareas de contexto corto (~ 150 palabras) con alta relevancia de respuesta (score: 0.813) y baja tasa de alucinación (score: 0.05)(Dharmawan & Witanti, 2025). Ollama se utiliza para el despliegue local, garantizando privacidad de datos, independencia de conexión a internet y eficiencia en el procesamiento de datos institucionales sensibles (Nussbaum et al., 2024).

Figura 4

Arquitectura del Sistema de Modelado RAG



Fuente: Los Autores

La arquitectura técnica del sistema está compuesta por un backed desarrollado con Flask un micro framework Open Source en Python que permite el desarrollo de aplicaciones web, caracterizado por si amplitud, potencia y flexibilidad al poder extenderse con módulos adicionales (Barros Manzano, 2022), que gestiona las APIs Restful, la integración con

MongoDB y el procesamiento asíncrono de documentos. El frontend, construido con Next.js, permite una experiencia fluida para la carga de documentos y la visualización de respuestas, se eligió este framework porque su renderizado del lado del servidor (SSR) mejora significativamente la velocidad de carga inicial y la experiencia del usuario, al entregar paginas listas para interactuar de forma inmediata (Srivastava et al., 2024).

Los componentes de modelado, se implementa un sistema de embeddings y búsqueda semántica que utiliza el modelo monic-embed-text para generar representaciones vectoriales del conocimiento técnico almacenado, estos vectores se almacenan y gestionan mediante FAISS lo que permite una recuperación eficiente por similitud. Además, se aplica normalización de los vectores para mejorar la precisión de búsqueda, tanto la pregunta original como los datos relacionados se envían al módulo RAG, donde el modelo Llama 3.2 genera una respuesta contextualizada.

La administración del contexto del diálogo la realizamos a través del almacenamiento del historial de MongoDB, el cual considera las interacciones, además de guardar los identificadores de usuario y las marcas de tiempo. El sistema, para que el contexto sea relevante, limita un historial a los últimos 8 mensajes, garantizando que el sistema solamente utilice los mensajes más recientes y relevantes.

Evaluación

La fase de evaluación consiste en corroborar los resultados obtenidos a los objetivos de negocio definidos inicialmente. Durante esta etapa se interpreta los resultados, se revisa el proceso que se ha seguido y se determina el contenido de las acciones posteriores que hay que seguir para verificar que el modelo cumple las expectativas que había (Schröer et al., 2021). En esta fase se hace la evaluación de la calidad mediante la norma ISO/IEC 9126, donde únicamente se toman en consideración las características de usabilidad, eficiencia y funcionalidad. Para ello se emplearán métricas de evaluación de tipo cualitativas (encuestas, cuestionarios...) y métricas de evaluación de tipo cuantitativas (tiempos de respuestas, tasas de acierto) basadas en estos factores solamente, dirigidas a usuarios, que son expertos en software e inteligencia artificial.

En este contexto, funcionalidad se refiere a la educación del sistema a los requisitos y necesidades, eficiencia a la relación entre el rendimiento del sistema y los recursos consumidos, y usabilidad al esfuerzo y facilidad con que los usuarios pueden utilizar el sistema (Uceda Martos et al., 2024), entre estas tres dimensiones permiten cuantificar

objetivamente la calidad del sistema y compararla con estándares de referencia.

Para aplicar estas características, se elaboraron tres instrumentos de evaluación en formato tabla, organizadas según las dimensiones mencionadas. Cada ítem será evaluado en una escala de cinco niveles: Excelente (E), Muy bueno (MB), Bueno (B), Regular (R) y Malo (M), permitiendo medir tanto la calidad técnica como la percepción del usuario.

Tabla 1

Instrumento de Evaluación - Funcionalidad

Ítem	Criterio de Evaluación
1	El chatbot responde correctamente a las preguntas relacionadas con los documentos técnicos
2	Las respuestas generadas por el sistema se ajustan a las necesidades del personal técnico
3	El sistema garantiza que la información facilitada esté segura y sea confidencial
4	Las respuestas permiten resolver problemas reales del área técnica
5	El sistema mantiene su funcionalidad en todos los módulos/documentos digitales cargados

Fuente: Los autores

Tabla 2

Tabla 1 Instrumento de Evaluación - Eficiencia

Ítem	Criterio de Evaluación
1	El tiempo de respuestas del chatbot es adecuado para las necesidades de consulta técnica
2	El sistema consume una cantidad razonable de recursos (CPU/RAM) durante el uso
3	Estabilidad del rendimiento durante periodos de uso continuo
4	El tiempo estimado en el uso de recursos permite atender múltiples consultas sin degradación
5	El tiempo de procesamiento de datos al subir nuevos documentos es aceptable

Fuente: Los autores

Tabla 3
Instrumento de Evaluación - Usabilidad

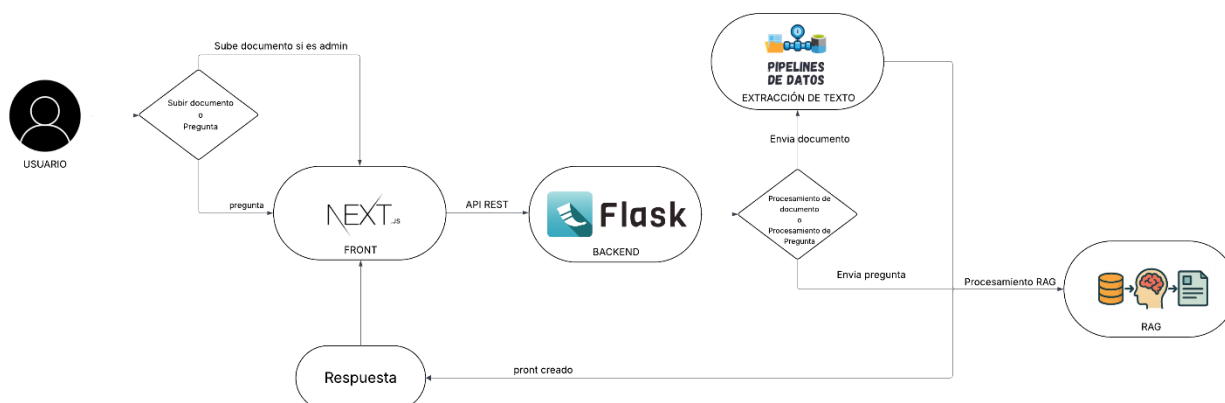
Ítem	Criterio de Evaluación
1	Es fácil aprender a utilizar el chatbot sin necesidad de capacitación extensa
2	La interfaz del chatbot es clara y fácil de usar
3	Es sencillo capacitar a otros usuarios para usar la herramienta
4	El usuario puede controlar la consulta o interacción de forma intuitiva
5	Los conceptos que maneja el sistema (mensajes, botones) son comprensibles para el usuario

Fuente: Los autores

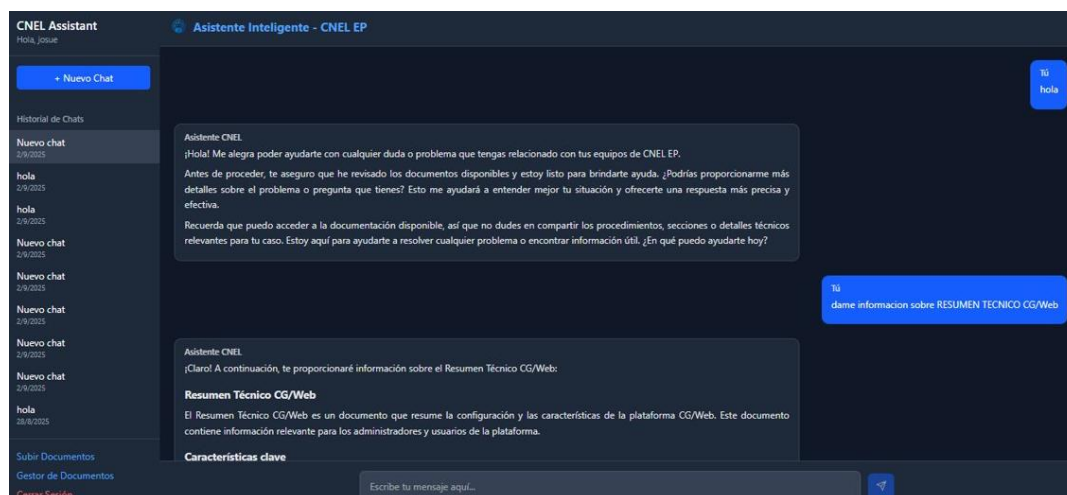
Despliegue

La etapa de despliegue representa la fase culminante del proceso, donde el conocimiento extraído debe organizarse y presentarse de manera clara y precisa para el cliente, permitiendo su aplicación efectiva; este proceso puede variar desde una simple generación de un informe hasta la compleja implementación de modelos para que los usuarios trabajen con nuevos datos (Tamariz Milla et al., 2025).

El sistema fue implementado en un entorno local utilizando una arquitectura RAG siendo full-stack con flask como backend y Next.js como fronted. La integración de todos los módulos sigue el flujo ilustrado de la Figura 5, garantizando una experiencia de usuario cohesiva y segura. El despliegue local con Ollama asegura que toda la información sensible de CNEL EP permanezca dentro del perímetro de la infraestructura institucional, cumpliendo con los protocolos de seguridad de la organización.

Figura 5*Integración total de todos los módulos de la aplicación***Fuente:** Los autores

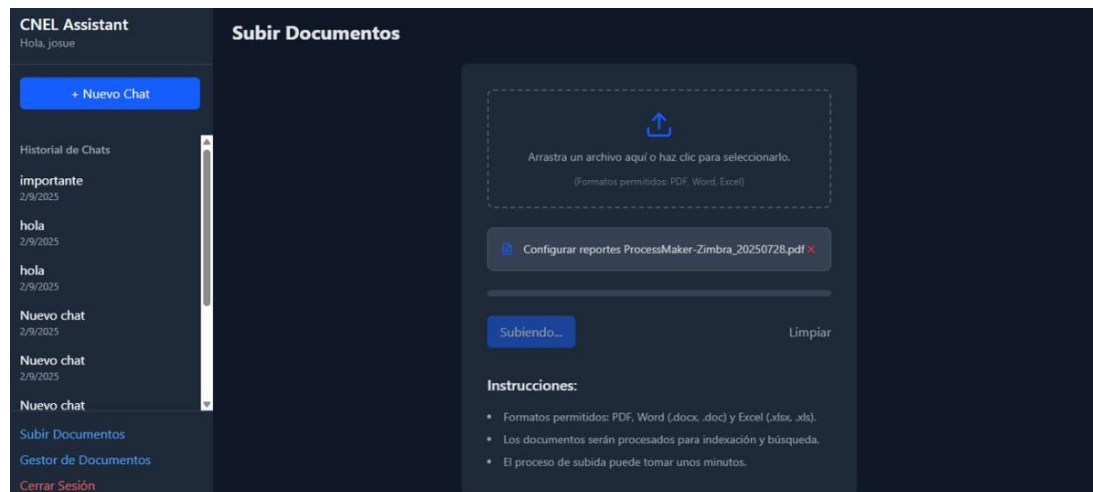
Con el objetivo de ilustrar el funcionamiento del prototipo en situaciones de consultas para la institución del CNEL EP, se presentan a continuación ejemplos de interacciones en la aplicación que demuestran la capacidad operativa. En este sentido, se pueden observar distintas interacciones simuladas, a partir de documentos técnicos reales de la empresa, que permiten ver la forma en la que el sistema entiende preguntas, obtiene información relevante de su base de conocimiento vectorial y responde de forma precisa y ajustada a situación.

Figura 6*Panel de interacción con el asistente virtual***Fuente:** Los autores

La Figura 6 representa la interfaz de usuario de la aplicación en la que el usuario puede redactar una pregunta sobre los documentos institucionales de CNEL EP. Esta pantalla muestra la forma en que los usuarios llevan a cabo las consultas utilizando IA.

Figura 7

Subida de documentos



Fuente: Los autores

La sección de la aplicación que puede cargar archivos o documentos en formato PDF, Word, Excel, que serán tratados y utilizados por el sistema RAG para entrenar la base de conocimiento y mejorar las respuestas que pueda ofrecer el sistema está presentada en la Figura 7.

RESULTADOS

La valoración de la aplicación web basada en la arquitectura RAG y modelo Llama 3.2 implementación en Ollama fue llevado a cabo bajo los criterios de funcionalidad, eficiencia y usabilidad, de acuerdo a la norma ISO/IEC 9126. A continuación, se muestran los principales hallazgos.

Si se consideran solo las funcionalidades, el sistema se llevó una puntuación de 4.5/5; lo que demuestra que el chatbot no solo responde de forma correcta a las preguntas técnicas, sino que también proporciona información contextualizada procedente de la institución en los documentos disponibles. Esto, a su vez, muestra que la arquitectura RAG es la adecuada para

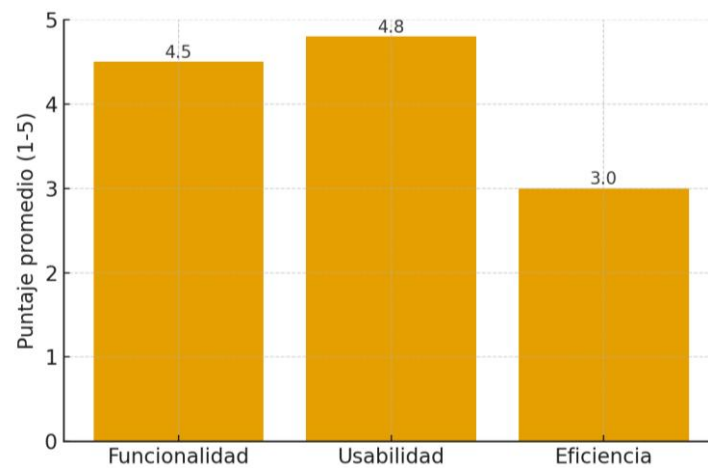
centralizar el conocimiento que está disperso en la CNEL EP.

En lo que se refiere a la usabilidad, la interfaz fue evaluada por los miembros participantes con una puntuación de 4.8/5, siendo calificada como clara e intuitiva. Además, gracias al diseño en Next.js, se facilitó la interacción del sistema y, por lo tanto, fue muy reducida la carga de capacitación anterior, convirtiéndose en una herramienta de fácil acceso para el personal técnico.

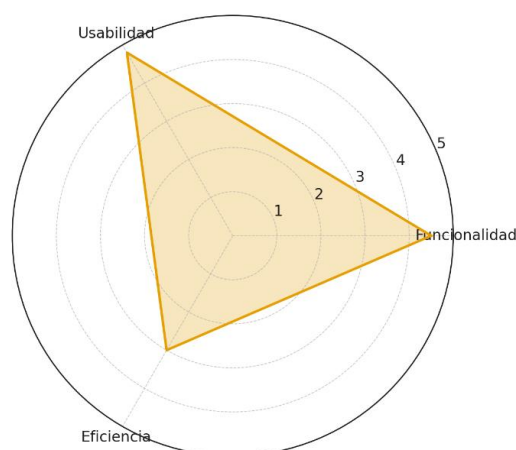
A pesar de que el criterio de eficiencia es también un aspecto importante, este número es definitivamente el más bajo comparado con las otras métricas, obteniendo un valor medio de 3.0/5. Los tiempos de respuesta resultantes fueron lentos a veces de la forma moderados, presentando un tiempo medio de respuesta en el orden de los 8.38 segundos por cada consulta a la base local, en Ollama con el modelo Llama 3.2. A pesar de que esta forma de trabajo asegura la privacidad de los datos manejados, interviene conscientemente en la velocidad de las interacciones y muestra un consumo de recursos computacionales demasiado alto en escenarios de uso intensivo.

Durante las pruebas iniciales, el sistema se ejecutó en un equipo de recursos de gama media con procesador 12th Gen Intel(R) Core(TM) i3-1215U (1.20 GHz), 16 GB de memoria RAM, disco SSD de 256 GB y un sistema operativo de 64 bits. Los resultados demostraron un consumo promedio de 2,5 a 3,2 GB de RAM durante las consultas y un uso de CPU de entre 35% y 55 % en picos de carga. El almacenamiento requerido por los documentos procesados e indexados fue aproximadamente 2,96 MB para un total de 34 documentos ($\approx$ 89KB por documento en promedio). Estos resultados confirman que el sistema puede operar de manera estable en hardware de gama media, lo que hace viable para la infraestructura actual de CNEL EP.

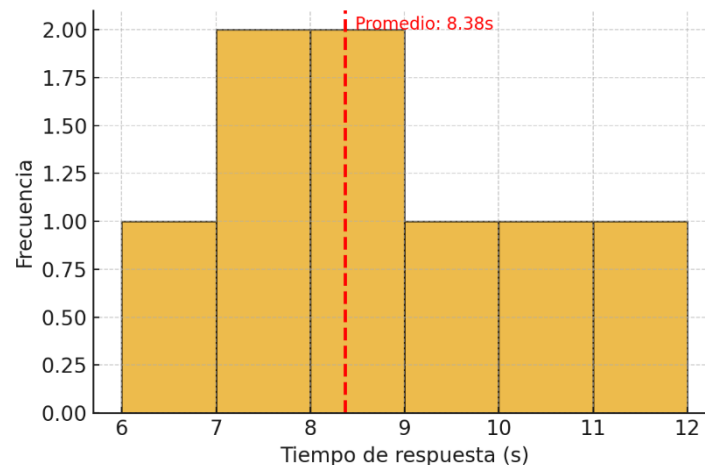
La Figura 8 muestra la comparación de puntajes promedio obtenidos en cada dimensión, destacando que la funcionalidad y la usabilidad superan la eficiencia.

Figura 8*Evaluación del sistema por dimensiones (ISO/IEC 9126)***Fuente:** Los autores

La Figura 9 presenta el radar de chart, donde se observa gráficamente el comportamiento balanceado del sistema entre funcionalidad y usabilidad, mientras que la eficiencia presenta una reducción significativa en comparación.

Figura 9*Radar de evaluación del sistema***Fuente:** Los autores (2025).

Por último, la Figura 10 ilustra la distribución de tiempos de respuesta medidos en segundos. La mayoría en un rango de 6 y 12 segundos, confirmando que la principal limitación del sistema radica en la velocidad de generación de respuestas.

Figura 10*Distribución de tiempos de respuesta del chatbot (Llama 3.2 en Ollama)***Fuente:** Los autores (2025).

Los resultados evidencian que la solución propuesta cumple de manera sobresaliente en funcionalidad y usabilidad, aunque requiere optimizaciones en eficiencia para mejorar la experiencia del usuario en el área de soporte técnico en CNEL EP.

DISCUSIÓN

La arquitectura RAG integra consultas a base de conocimiento institucionales con modelos generativos, lo que tiende a superar a chatbots tradicionales basados únicamente en LLM. Estudios recientes confirman esta ventaja: por ejemplo, (Steybe et al., 2025) evaluaron un chatbot clínico “GuideGPT” usando RAG y GPT-4, obteniendo puntuaciones significativamente mejores en contenido y precisión científica que en GPT-4 generativo. De modo similar, (Soudani et al., 2024) compararon RAG vs. Ajuste fino en dominios con información poco frecuente y hallaron que RAG “consistentemente supera” al ajuste fino en modelos de menor tamaño, especialmente al combinarlo con refinamientos adicionales. En la práctica de soporte técnico, esto sugiere que nuestro sistema RAG puede mejorar consultas especializadas y actualizaciones de conocimiento sin requerir reentrenamientos constantes, mitigando alucinaciones y ofreciendo trazabilidad a fuentes confiables. No obstante, implementar RAG implica mayor complejidad de arquitectura (índices vectoriales, fragmentación de documentos, pipelines de recuperación), un costo computacional adicional y dependencia de la calidad de la base de datos vectorial.

El modelo Llama 3.2 (3B) fue elegido por su balance entre capacidad y eficiencia, estudios de meta demuestran que la familia Llama 3 entrega una calidad comparable a modelos líderes como GPT-4 en múltiples tareas (Grattafiori et al., 2024). Efectivamente (Dharmawan & Witanti, 2025) hallaron que Llama 3.2 logró alta relevancia de respuestas (0.813) y mínima alucinación (0.05) en generación de preguntas a partir del texto, estos valores refuerzan su aptitud para consultas breves y técnicas como las de nuestro dominio. Es importante notar que, aunque los LLM propietarios (p.ej. GPT-4) aún sobresalen ligeramente en desempeño, los modelos de código abierto como Llama 3.2 cierran brecha permitiendo despliegues locales seguros (Amugongo et al., 2025).

El sistema se ejecuta localmente para preservar la confidencialidad de la información institucional, maximiza el control de datos y la privacidad, ya que toda la información permanece dentro de la infraestructura de CNEL EP. La necesidad de operar “on-premise” fue señalada en casos reales similares, una empresa del sector (SAP en Oil&Gas) optó por un chatbot local precisamente para garantizar privacidad y control total de los datos (Alali et al., 2025). Este enfoque evita depender de conexiones externas y asegura que ningún dato sensible salga de la infraestructura de la institución. Sin embargo, el despliegue local implica desafíos como la inferencia de modelos como Llama en hardware limitado introduce latencias notables (nuestro promedio fue de 8.38 segundos por consulta) y medio uso de CPU/RAM. Además, estudios recientes advierten que marcos de despliegue locales (p.ej. Ollama) suelen venir con configuraciones inseguras por defecto (Hou et al., 2025), Ollama usando para ejecutar Llam localmente, expone APIs sin autenticación en muchos casos, lo que permitiría operaciones no autorizadas (borrado de modelos, ejecución remota de código, etc.), por lo tanto, se requieren prácticas de seguridad más estrictas (firewalls, credenciales, TSL) en entornos locales. En contraste, las soluciones en la nube ofrecen generalmente menores tiempos de respuesta y escalabilidad automática, pero a costa de costos recurrentes, dependencia de internet y potenciales caídas del servicio (algo documentado en proveedores locales) (Jang & Morabito, 2025). Nuestro despliegue local otorga máxima privacidad y control (cumpliendo los requerimientos organizacionales), mientras en la investigación indica que una desventaja inherente es la eficiencia reducida y la necesidad de reforzar la seguridad operativa.

CONCLUSIONES

El trabajo de investigación del desarrollo de la aplicación web con arquitectura RAG y el modelo Llama 3.2 demostró que es posible centralizar eficazmente la información técnica dispersa de CNEL EP, ofreciendo un mecanismo confiable y accesible para la generación de respuestas contextualizadas a consultas de soporte técnico. De los resultados de la evaluación correspondiendo a la norma ISO/IEC 9126, se puede confirmar que la herramienta cumple de forma excepcional con el objetivo principal, alcanzando niveles altos en funcionalidad (4.5/5) y en usabilidad (4.8/5).

Es pertinente indicar que la solución no es un sustituto de la capacidad de análisis que posea el personal técnico, sino que se convierte en una ayuda inteligente que optimiza el acceso al conocimiento institucional al validar la precisión de la información recuperada y prevenir la inconsistencia a la hora de consultar los documentos. Esto facilitará que el usuario entienda y aplique con mayor seguridad los conceptos técnicos dedicados, redoblando la curva de aprendizaje y el número de errores por desinformación.

La arquitectura RAG cuando se utiliza con tecnologías como Flask, Next.js, FAISS y el despliegue local con Ollama, puede permitir implementar exitosamente un sistema de gestión del conocimiento que permite tener bajo control la privacidad de los datos y que se adapta a la infraestructura existente de CNEL EP. Sin embargo, el despliegue local presentó muchos problemas de eficiencia, ya que las medias de los tiempos de respuesta están por encima de los 8.38 segundos, lo que sigue indicando que es un aspecto esencial a mejorar para el futuro. En proyección, se propone optimizar el rendimiento con la aplicación de técnicas de cuantización del modelo, escalando en el índice vectorial o, potencialmente, migrando a entornos híbridos para asegurar la seguridad, pero mejorar los tiempos de respuesta. A su vez, se sugiere validar la herramienta con un grupo de usuarios más amplio con diferentes niveles de experiencia, así como estudiar la inclusión de nuevos formatos documentales y protocolos de seguridad reforzado para con un tipo específico de contenido.

De tal manera que, el presente estudio puede establecer las bases, tanto metodológicas como técnicas, para un sistema de gestión del conocimiento, siendo este sistema validado inicialmente y arrojando un, sin lugar a dudas, alto potencial para mejorar procesos en CNEL EP e implicar el uso de inteligencia artificial en el sector eléctrico ecuatoriano.

REFERENCIAS BIBLIOGRÁFICAS

- Ahumada, H. C., Póliche, M. V., & Contreras, N. A. (2020). Chatbot asistente para atención de aspirantes en la Facultad de Tecnología y Ciencias Aplicadas de la UNCA. *Jornadas Argentinas de Informática*, 49, 142-152. <https://49jaiio.sadio.org.ar/pdfs/asai/ASAI-11.pdf>
- Alali, F., Bashar, A., Aldawsari, H., & Mahmood, S. (2025). Comparative Analysis of Industrial SAP Chatbots: RAG-LLM and Cloud-based Approaches. *Proceedings of the International Conference on Social Sciences and Business (ICSSB 2025)*, 1, 1-8. https://faculty.pmu.edu.sa/Admin/ckEditor_Uploads/133969093386114239_Full_paper_ICSSB25_draft.pdf
- Alghanemi, J., & Al Mubarak, M. (2022). The Role of Artificial Intelligence in Knowledge Management. En A. Hamdan, A. Harraf, P. Arora, B. Alareeni, & R. K. Hamdan (Eds.), *Studies in Computational Intelligence* (pp. 359-373). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-030-99000-8_20
- Amugongo, L. M., Mascheroni, P., Brooks, S., Doering, S., & Seidel, J. (2025). Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4, 1-33. <https://doi.org/https://doi.org/10.1371/journal.pdig.0000877>
- Barros Manzano, M. V. (2022). Asistente conversacional (chatbot) basado en un sistema de gestión de conocimientos para atención a los clientes de la empresa aceros industriales hgb. [Tesis de pregrado, Universidad técnica de Ambato]. Archivo digital. <https://repositorio.uta.edu.ec/handle/123456789/34351>
- Bregón Bregón, A., & Silvestre Vilches, J. (2021). Implementación de una herramienta basada en PLN para la detección y anonimización de datos personales en documentos [Tesis de Maestría, Universidades de Burgos, León y Valladolid]. Archivo digital. <https://uvadoc.uva.es/handle/10324/49995>
- Calatrava Hurtado, D. J. (2019). Desarrollo de una solución para la visualización y análisis de sentimiento en base a los mensajes que realizan los usuarios a una empresa mediante una red social [Tesis de pregrado, Universidad de las Américas]. Archivo digital. <http://dspace.udla.edu.ec/handle/33000/10600>
- CNEL EP. (2016). ¿Quiénes Somos? - CNEL EP. [cnelep.gob.ec. https://www.cnelep.gob.ec/quienes-somos/](https://www.cnelep.gob.ec/quienes-somos/)
- Dharmawan, T., & Witanti, A. (2025). Evaluasi llama3.2 3b untuk menghasilkan soal otomatis dengan deepeval berdasarkan metrik answer relevancy dan hallucination. *Jurnal*

- Informatika Teknologi dan Sains, 7, 242-248.
<https://doi.org/https://doi.org/10.51401/jinteks.v7i1.5423>
- Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., & Jégou, H. (2025). The Faiss library. arXiv preprint arXiv:2401.08281, 3, 1-24.
<https://doi.org/https://doi.org/10.48550/arXiv.2401.08281>
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., ... Ma, Z. (2024). The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783, 3, 1-92.
<https://doi.org/https://doi.org/10.48550/arXiv.2407.21783>
- HeidiSteen. (2025). RAG e inteligencia artificial generativa - Azure AI Search | Microsoft Learn. Microsoft.com. <https://learn.microsoft.com/es-es/azure/search/retrieval-augmented-generation-overview?tabs=docs>
- Hou, X., Han, J., Zhao, Y., & Wang, H. (2025). Unveiling the Landscape of LLM Deployment in the Wild: An Empirical Study. arXiv preprint arXiv:2505.02502, 1, 1-12.
<https://doi.org/https://doi.org/10.48550/arXiv.2505.02502>
- Jang, S., & Morabito, R. (2025). Edge-First Language Model Inference: Models, Metrics, and Tradeoffs. arXiv preprint arXiv:2505.16508, 1, 1-7.
<https://doi.org/https://doi.org/10.48550/arXiv.2505.16508>
- Maruri, B., Reina, R., Lara, O., Espinoza, &, Reinaldo, N., Gregorio, Á., Reina, R., Raúl, F., Paola, M., & Martínez, E. (2025). Automatización del soporte al cliente mediante un chatbot con IA integrado al CRM: caso de estudio empresa EDITRATECH. GADE: Revista Científica, 5, 206-219. <https://doi.org/10.63549/RG.V5I3.705>
- Nussbaum, Z., Duderstadt, B., & Mulyar, A. (2024). Nomic Embed Vision: Expanding the Latent Space. arXiv preprint arXiv:2406.18587, 1, 1-7.
<https://doi.org/https://doi.org/10.48550/arXiv.2406.18587>
- Sánchez García, S. (2024). Integración de arquitectura de generación aumentada por recuperación (RAG) en la plataforma Wattwin. Dipòsit Digital de Documents de la UAB, 1, 1-15.
<https://ddd.uab.cat/record/298985>
- Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. Procedia Computer Science, 181, 526-534.
<https://doi.org/10.1016/j.procs.2021.01.199>

- Skinner Barbosa, M. J., & García Díaz, J. F. (2025). Chatbot como herramienta de apoyo para mejorar los procesos de inserción laboral [Tesis de pregrado, Universidad Nacional Abierta y a Distancia UNAD]. Archivo digital. <https://repository.unad.edu.co/handle/10596/69161>
- Soudani, H., Kanoulas, E., & Hasibi, F. (2024). Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge. SIGIR-AP 2024 - Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, 12-22. <https://doi.org/10.1145/3673791.3698415>
- Srivastava, S., Shukla, H., Landge, N., Srivastava, A., & Jindal, D. (2024). A Comprehensive Review of Next.js Technology: Advancements, Features, and Applications. SSRN Electronic Journal, 1, 1-4. <https://doi.org/http://dx.doi.org/10.2139/ssrn.4831070>
- Steybe, D., Poxleitner, P., Aljohani, S., Herlofson, B. B., Nicolatou-Galitis, O., Patel, V., Fedele, S., Kwon, T. G., Fusco, V., Pichardo, S. E. C., Obermeier, K. T., Otto, S., Rau, A., & Russe, M. F. (2025). Evaluation of a context-aware chatbot using retrieval-augmented generation for answering clinical questions on medication-related osteonecrosis of the jaw. Journal of Cranio-Maxillofacial Surgery, 53, 355-360. <https://doi.org/10.1016/j.jcms.2024.12.009>
- Tamariz Milla, M. A., Avila Narvaez, M. A., Denegri Tineo, E. E., Cruz Solano, J. D., & Torres Alama, E. A. (2025). Factores de la atención de cobranza que conduce a la generación de un compromiso de pago [Tesis de Maestría, Universidad Peruana de Ciencias Aplicadas]. Archivo digital. <http://hdl.handle.net/10757/685067>
- Tamayo Cárdenas, Y. L., & Villavicencio Martínez, P. A. (2018). Módulo para el análisis de datos basado en CRISP-DM para la plataforma XEDRO-GESPRO [Tesis de pregrado, Universidad de las Ciencias Informáticas]. Archivo digital. <https://repositorio.uci.cu/jspui/handle/123456789/10216>
- Uceda Martos, P. J., Polo Chávarri, L. C., & Cruzado Asencio, G. A. (2024). La influencia de los chatbots en el aprendizaje remoto como una estrategia de autorregulación en tiempos de pandemia. En M. Fernández Ferrer (Ed.), Chatbots en Educación Tendencias actuales y desafíos futuros (Número 21, pp. 93-120). Ibero-American Law Institute.

©2026 por los autores. Este artículo es de acceso abierto y distribuido según los términos y condiciones de la licencia Creative Commons Atribución-NoComercial-CompartirIgual 4.0 Internacional (CC BY-NC-SA 4.0) (<https://creativecommons.org/licenses/by-nc-sa/4.0/>)