



Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

Trabajo fin de estudio presentado por:	Daniel Jeronimo Chica Quinchia Juan David García Ramírez
Tipo de trabajo:	TFM-Inteligencia Artificial
Director/a:	Deidy Gutiérrez Niño
Fecha:	26 marzo de 2026

- Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

Índice

1 Resumen	3
2 Abstract	3
3 Introducción	5
3.1 Objetivo General	5
3.2 Objetivos Específicos	5
3.3 Metodología	6
3.3.1 Identificación y Análisis del Problema	6
3.3.2 Planificación de la Solución	7
4 Identificación y descripción un problema al interior de la organización	7
4.1 Descripción de la empresa a intervenir	8
4.1.1 Misión	8
4.1.2 Visión	8
4.1.3 Mapa de procesos	9
4.1.4 Descripción del proceso a intervenir	10
4.1.5 Descripción del problema que presenta el proceso	10
Tabla 1	11
Tabla 2	11
Tabla 3	12
5 Establecimiento de planes de acción para la solución del problema	14
5.1 Establecimiento de objetivos	14
5.2 Descripción de la solución	15
5.3 Secuencia de actividades para ejecución de la solución	16
5.4 Descripción de recursos necesarios para la ejecución de la solución	17
6 Definición del alcance de la solución, partes interesadas, cronograma y presupuestos.	19
6.1 Alcance final de la solución	19
6.2 Identificación de responsables de las actividades para ejecución	19
6.3 Cronograma de ejecución	20
6.4 Presupuesto	21
7 Referencias bibliográficas	23

■ Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

1 Resumen

Este proyecto empleó la metodología de investigación-acción para enfrentar un problema frecuente en las entidades públicas: la dispersión de documentos, la falta de coherencia en las respuestas institucionales y los largos tiempos que toma acceder a información confiable. Para comprender la magnitud del reto, se realizó un diagnóstico que incluyó un análisis FODA, entrevistas, observación directa y la revisión de métricas operativas. Estos insumos permitieron identificar una fuerte dependencia de expertos y la ausencia de una herramienta que centralizará la búsqueda de información.

Con base en estos resultados, se diseñó un asistente de inteligencia artificial generativa desplegado bajo una arquitectura On-Premises. La solución utiliza modelos de lenguaje de gran escala y la técnica de Generación Aumentada por Recuperación (RAG), lo que permite ofrecer respuestas más precisas, verificables y alineadas con los documentos oficiales de la entidad.

Los primeros resultados del piloto muestran avances importantes: una reducción del 28 % en los tiempos de búsqueda, un incremento del 45 % en las respuestas con citación documental y una disminución del 30 % en la dependencia de expertos para resolver consultas. Estos logros evidencian el potencial de la herramienta para mejorar la eficiencia, fortalecer la trazabilidad de la información y apoyar una gestión del conocimiento más sólida y sostenible dentro del Ministerio.

Palabras clave: Gestión del conocimiento; inteligencia artificial generativa; LLM on-premises; RAG; sector público

2 Abstract

This project employed action research methodology to address a common problem in public entities: the dispersion of documents, the lack of consistency in institutional responses, and the lengthy time it takes to access reliable information. To understand the magnitude of the challenge, a diagnostic assessment was conducted, including a SWOT analysis, interviews, direct observation, and a review of operational metrics. This input revealed a strong reliance on experts and the absence of a tool to centralize information retrieval.

Based on these findings, a generative artificial intelligence assistant was designed and deployed on-premises. The solution utilizes large-scale language models and Retrieval Augmented Generation (RAG) techniques, enabling it to provide more accurate, verifiable answers aligned with the entity's official documents.

The initial results of the pilot project show significant progress: a 28% reduction in search times, a 45% increase in responses with document citations, and a 30% decrease in reliance on experts to resolve queries. These achievements demonstrate the tool's potential to improve efficiency, strengthen information traceability, and support more robust and sustainable knowledge management within the Ministry.

- Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

Keywords: Knowledge management; generative artificial intelligence; on-premises LLM; RAG; public sector

Instituto Europeo de Posgrado ©

- Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

3 Introducción

Este proyecto de aplicación en investigación-acción se centró en la identificación y resolución del problema de dispersión documental, altos tiempos en la consulta de información y la ausencia de un mecanismo centralizado de consulta inteligente dentro de las entidades públicas en Colombia (Ministerios), lo que afecta la eficiencia operativa y la capacidad de ofrecer información confiable tanto a las áreas internas dentro de la entidad como a la ciudadanía en general, permitiendo integrar teoría y práctica en un entorno real. Mediante un enfoque de investigación-acción, se desarrollaron habilidades de diagnóstico, análisis y planificación orientadas a mejorar la eficiencia y efectividad de un proceso crítico en la organización. La estructura del proyecto incluyó una revisión exhaustiva del contexto organizacional y la implementación de una solución factible que responde a las necesidades de la organización y optimiza su desempeño.

El proyecto sienta bases en fuentes de renombre como Huang et al. (2025), la cual demuestra que el despliegue on-premises de los LLMs mejora la privacidad de los datos y reduce riesgos de filtración y extracción del modelo, lo que respalda la decisión de implementar un asistente conversacional local en lugar de APIs públicas en una entidad gubernamental con información sensible. También Tabassi (2023), mediante el NIST AI RMF 1.0, nos aporta un marco operativo de gestión de riesgos de IA (gobernar-mapear-medir-gestionar) que utilizamos para estructurar la gobernanza del piloto, definir roles, controles y métricas de desempeño y de confianza del sistema y por último Gao et al. (2024) sintetiza la evidencia sobre Retrieval-Augmented Generation (RAG) y muestran que este patrón reduce alucinaciones y ancla las respuestas en documentos citables, lo cual justifica el uso de RAG para conectar los LLMs con el repositorio normativo y procedimental de la entidad, garantizando respuestas precisas y trazables.

3.1 Objetivo General

El objetivo principal del proyecto fue desarrollar una solución operativa al problema de dispersión documental, con altos tiempos de respuesta y ausencia de un mecanismo inteligente de consulta institucional, aprovechando la información que ha sido recolectada por años por estas entidades públicas (Ministerios) específicamente en Colombia, mediante el diseño e implementación de un asistente conversacional tipo ChatGPT On-Premises o Local. La intervención se orientó no sólo a corregir el problema identificado, sino también a proporcionar una propuesta sostenible que pudiera integrarse a la estructura de la organización, elevando los niveles de eficiencia y alcanzando los estándares de desempeño requeridos en el proceso intervenido.

3.2 Objetivos Específicos

Para cumplir con el objetivo general, se establecieron tres objetivos específicos:

■ Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

1. Realizar una identificación detallada del problema en su contexto organizacional, determinando causas raíz y evaluando su impacto en el desempeño general del proceso.
2. Diseñar un plan de acción estructurado con objetivos claros, actividades específicas y recursos definidos para la ejecución de la solución.
3. Establecer el alcance de la intervención, definiendo responsables, cronograma y presupuesto, para asegurar que la implementación sea organizada y efectiva.

3.3 Metodología

El proyecto se desarrolló mediante la metodología de investigación-acción, la cual facilitó la colaboración continua entre el equipo de estudiantes y docente asesor. Esta metodología permitió un enfoque sistemático para identificar problemas, formular hipótesis y aplicar soluciones en un contexto real, ajustando las estrategias conforme se avanzaba en el proyecto. La investigación-acción proporcionó un marco flexible y participativo para ejecutar el proyecto de manera estructurada, generando una intervención bien fundamentada y con alto potencial de impacto.

El desarrollo del proyecto se estructuró en tres etapas secuenciales, cada una de las cuales abordó un aspecto específico de la intervención. En la primera etapa, se identificó y analizó el problema organizacional, incluyendo una descripción detallada de la organización, su misión, visión y el proceso específico afectado. La segunda etapa se centró en la planificación de la solución, estableciendo objetivos y actividades concretas para abordar el problema. En la tercera fase, se definieron aspectos operativos como el cronograma, el presupuesto y la asignación de roles.

3.3.1 Identificación y Análisis del Problema

En la primera etapa, se realizó un análisis exhaustivo de la organización y del contexto en el cual se desarrolla el proceso afectado. Se documentó el problema identificando sus causas raíz, efectos y puntos críticos que afectaban el desempeño del proceso. Este diagnóstico inicial fue esencial para estructurar la solución y asegurar que la intervención respondiera de manera precisa a las necesidades de la organización.

Para el diagnóstico y análisis de la problemática, se aplicó un enfoque de investigación-acción que combinó técnicas cualitativas y cuantitativas. En una primera fase, se realizó un análisis FODA orientado a identificar debilidades en la gestión del conocimiento institucional, especialmente la dispersión documental y en la ausencia de herramientas inteligentes de búsqueda. Posteriormente, se desarrollaron entrevistas con líderes de cada uno de los procesos y el personal técnico para comprender sus necesidades de análisis de la información y las dificultades en la consulta de normativas y procedimientos. Este proceso culminó con la observación directa de los flujos de trabajo actuales, lo que permitió evidenciar la dependencia de expertos, la duplicidad de búsquedas y los altos tiempos de respuesta y análisis de la información. El diagnóstico integral nos proporcionó la evidencia necesaria para justificar una solución basada en un asistente conversacional ChatGPT

- Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

on-premises o local capaz de integrar, unificar, garantizar la seguridad de la información y recuperar información de forma precisa.

3.3.2 Planificación de la Solución

En la segunda etapa, se desarrolló un plan de acción detallado que incluyó la definición de objetivos específicos y una secuencia de actividades que permitiría mitigar o resolver el problema identificado. Esta fase incluyó el diseño de una estrategia de ejecución organizada y estructurada que facilitará el logro de los objetivos propuestos y maximizará la eficiencia del proceso intervenido. La planificación detallada de actividades y recursos fue clave para asegurar una implementación precisa y efectiva.

3.3.2.1 Definición de Alcance y Recursos

La tercera etapa se enfocó en delimitar el alcance de la intervención y los recursos necesarios. Se establecieron responsables para cada actividad, y se elaboró un cronograma de ejecución para controlar el avance de la implementación. Asimismo, se desarrolló un presupuesto que detalla los costos asociados a cada fase, permitiendo optimizar la asignación de recursos. Esta fase aseguró que todas las variables críticas estuvieran alineadas para una ejecución sin interrupciones.

- Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

4 Identificación y descripción un problema al interior de la organización

En esta sección se identifica y describe con claridad el problema asociado a la dispersión documental, los tiempos de respuesta extensos y la ausencia de un mecanismo inteligente de consulta dentro del proceso de acceso a la información institucional de la entidad. El análisis parte de una breve descripción del contexto organizacional y del área donde se manifiesta la problemática, evidenciando la fragmentación de repositorios (Silos de Información/Falta de Gobierno), la dependencia de expertos y la falta de estandarización en la búsqueda de información. Se examinan las causas raíz y los efectos del problema, destacando su impacto en la eficiencia operativa, la calidad de las respuestas y la capacidad de la organización para ofrecer información precisa, coherente y oportuna. Asimismo, se incorporan datos y evidencias recabadas durante el diagnóstico que permiten sustentar la necesidad de una solución a esta problemática. Esta descripción detalla la necesidad de una solución del tipo asistente conversacional ChatGPT On-Premises o local como solución tecnológica para centralizar, agilizar y fortalecer el acceso seguro al conocimiento institucional y al final que permita prestar un mejor servicio a la ciudadanía.

4.1 Descripción de la empresa a intervenir

El Ministerio es un organismo público del orden nacional encargado de formular, orientar y coordinar políticas públicas en su sector. Su estructura institucional integra áreas estratégicas, misionales y de apoyo que gestionan grandes volúmenes de información normativa, técnica y administrativa necesarios para la toma de decisiones y la atención de requerimientos ciudadanos. La entidad enfrenta desafíos relacionados con la dispersión de documentos, la dependencia de expertos para consultas especializadas y la necesidad de fortalecer sus procesos de gestión del conocimiento. En este contexto, la implementación de soluciones tecnológicas, como un asistente conversacional representa una oportunidad para mejorar la eficiencia operativa, garantizar el acceso oportuno a la información y consolidar prácticas de gobierno digital alineadas con los estándares del sector público colombiano.

4.1.1 Misión

El Ministerio tiene como misión formular, coordinar y ejecutar políticas públicas orientadas al fortalecimiento institucional, la mejora de los servicios a la ciudadanía y la gestión eficiente del conocimiento público. A través de una administración moderna, segura y transparente, la entidad promueve el acceso oportuno a la información, garantizando la protección de los datos, la continuidad operativa y la adopción de soluciones tecnológicas que apoyen la toma de decisiones y optimicen los procesos internos.

4.1.2 Visión

El Ministerio aspira a consolidarse como un referente nacional en innovación pública, fortaleciendo su capacidad para ofrecer servicios oportunos, confiables y centrados en las necesidades de la ciudadanía. Proyecta un modelo de gestión institucional basado en el uso seguro, ético y responsable de tecnologías avanzadas, que permitan integrar y democratizar

■ Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

el acceso al conocimiento público. Así, la entidad visualiza un futuro en el que soluciones como plataformas de inteligencia artificial desplegadas de manera local incluyendo asistentes conversacionales impulsen la eficiencia operativa, garanticen la protección de los datos y contribuyan a una administración pública más transparente, moderna y efectiva.

4.1.3 Mapa de procesos

El mapa de procesos del Ministerio se organiza en tres grandes categorías: procesos estratégicos, procesos misionales y procesos de apoyo. En conjunto, estos procesos articulan el funcionamiento institucional y permiten visualizar de manera integral cómo fluye la información y cómo se toman las decisiones y ejecución de las funciones propias del Ministerio.

En la parte superior se ubican los procesos estratégicos, responsables de la formulación de políticas, la planeación institucional, la gestión del riesgo y la toma de decisiones de alto nivel. Estos procesos orientan el rumbo de la entidad y determinan los lineamientos que impactan directamente la gestión del conocimiento y la disponibilidad de información.

En el núcleo del mapa se encuentran los procesos misionales, donde se agrupan las actividades esenciales relacionadas con el cumplimiento de la función pública asignada al Ministerio. Incluyen la formulación y seguimiento de políticas sectoriales, la coordinación con otras entidades del estado (Interoperabilidad), la generación y divulgación de información oficial y la atención de requerimientos ciudadanos. Es en este nivel donde se evidencia con mayor fuerza la necesidad de contar con sistemas de consulta eficientes y confiables, considerando el volumen de normativas, lineamientos y documentos técnicos que se producen y gestionan diariamente.

Finalmente, los procesos de apoyo proporcionan los servicios internos necesarios para que la entidad opere de manera eficiente, como tecnologías de la información, gestión documental, talento humano, comunicaciones y jurídica. En estos procesos se identifican las principales brechas relacionadas con la dispersión documental, la duplicidad de repositorios, los tiempos prolongados de búsqueda de información y la ausencia de herramientas inteligentes de consulta, elementos que fundamentan la pertinencia de implementar una solución como ChatGPT On-Premises o local para optimizar el acceso al conocimiento institucional.

Ilustración 1. Mapa de procesos

- Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

MAPA DE PROCESOS



Fuente: Elaboración propia mediante Copilot.

4.1.4 Descripción del proceso a intervenir

El proceso específico a intervenir es el de Consulta Institucional de Información y Gestión del Conocimiento utilizado por las áreas misionales y la alta dirección. Actualmente, la atención de consultas se realiza de forma manual y dispersa en múltiples repositorios, lo que obliga a escalar preguntas a expertos y prolonga los tiempos de respuesta. La ausencia de un mecanismo de búsqueda semántica con trazabilidad de fuentes limita la consistencia de las respuestas y dificulta la toma de decisiones en eventos críticos. La intervención propone incorporar un asistente conversacional ChatGPT On-Premises con RAG para unificar la consulta, citar la fuente oficial y garantizar seguridad y soberanía de los datos.

4.1.5 Descripción del problema que presenta el proceso

Se utilizó un enfoque de investigación-acción combinando entrevistas con directivos, líderes de proceso y analistas. Análisis de flujos de consulta para establecer la línea base (tiempos de respuesta, dependencia de los expertos) y el Diagrama de Causa-Efecto (Ishikawa) para mapear las causas raíz. Los hallazgos se consolidaron en una Matriz DOFA, priorizando debilidades y riesgos que impactan la oportunidad y trazabilidad de la información. Con esta información se definieron los requisitos (RAG, citación verificable, control de acceso) y las métricas de éxito para la solución con ChatGPT On-Premises.

Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

Tabla 1

Principales Hallazgos de Entrevistas (n = 12)

Código	Hallazgo (síntesis)	Evidencia cualitativa (cita breve)	Implicación para la solución
E-01	Dependencia de expertos para localizar/interpretar normativa	"Una respuesta requiere validar con 2-3 personas antes de enviarla."	RAG con citas verificables
E-02	Dispersión de repositorios y versiones	"Está en intranet, SharePoint y GED."	Índice único + metadatos
E-03	Tiempos de respuesta altos en picos	"Con una circular nueva, se congestiona todo."	Asistente on-prem 24/7
E-04	Inconsistencia en respuestas	"Responden distinto según quién atiende."	Plantillas guardrails +
E-05	Riesgos al compartir extractos	"Se envían PDFs completos por correo."	Acceso documento por auditoría +

Nota: Elaboración propia a partir de entrevistas con directivos, líderes y analistas.

Tabla 2

Métricas del Flujo "As-Is" (Muestra: 120 Casos, 6 Semanas)

Métrica	Valor línea base	Observación
---------	------------------	-------------

■ Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

Tiempo medio de respuesta (TMO)	22 h 18 min	Búsqueda manual y validaciones en serie
Contacto por caso (prom.)	1,7	Falta de trazabilidad
Consultas con cita documental (%)	42	Poca referencia a versión vigente
Dependientes de experto (%)	68	Cuello: disponibilidad del experto
Incidencias por documento desactualizado (%)	19	Problemas de versionamiento

Nota. Fuente: Elaboración propia con registro de consultas internas (6 semanas). Objetivos del piloto: reducir TMO $\geq 30\%$, elevar porcentaje de respuestas con cita $\geq 80\%$ y reducir dependencia de experto $\leq 35\%$.

Tabla 3

Matriz DOFA del Proceso de Consulta Institucional

Fortalezas (F)	Oportunidades (O)
F1. Acervo documental digitalizado	O1. LLM on-prem y vector stores
F2. Expertos temáticos disponibles	O2. Lineamientos de Gobierno Digital

■ Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

F3. Procedimientos formales	O3. Apertura a pilotos de IA responsable
Debilidades (D)	Amenazas (A)
D1. Dispersión de repositorios y metadatos	A1. Riesgo de divulgar datos
D2. Tiempos altos por búsqueda manual	A2. Picos de demanda normativa
D3. Inconsistencias y baja trazabilidad	A3. Dependencia de proveedor externo

Nota. Elaboración propia. Estrategias: FO (RAG on-prem), DO (normalización de metadatos), FA (controles y auditoría), DA (gestión del cambio).

- Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

5 Establecimiento de planes de acción para la solución del problema

En esta segunda etapa del proyecto se define la ruta de ejecución que permitirá transformar la gestión del conocimiento dentro del Ministerio mediante la integración de modelos de lenguaje de gran escala (LLM) desplegados de forma local (Instalaciones del cliente). Siguiendo el enfoque de investigación-acción, esta fase convierte los hallazgos del diagnóstico en intervenciones concretas orientadas a reducir la dispersión documental, disminuir los tiempos de respuesta institucional y garantizar la coherencia y trazabilidad en la consulta de información especializada (Kemmis & McTaggart, 2005).

El plan de acción integra simultáneamente consideraciones técnicas, organizacionales y de gobernanza, articulando el despliegue de un asistente de IA generativa bajo una arquitectura On-Premises. Este esquema permite fortalecer los procesos de gestión del conocimiento, asegurar la protección de datos sensibles (bastante importante para los Ministerios en general) y mantener el cumplimiento de estándares normativos aplicables al sector público. Asimismo, este enfoque no solo propone una solución tecnológica, sino que establece un componente estratégico dentro del Ministerio que habilita la sostenibilidad del proyecto en el entorno institucional, coherente con la naturaleza cíclica de la investigación-acción y su principio de mejora continua (Lewin, 1946; Kemmis & McTaggart, 2005).

5.1 Establecimiento de objetivos

La definición de los objetivos de esta propuesta nace de la necesidad de mejorar la manera en que se accede a la información dentro del Ministerio. Hoy en día, consultar documentos es un proceso lento, poco práctico y en ocasiones inseguro. Además, gran parte de esta información está distribuida en distintos lugares y no siempre es fácil de encontrar. Con esta propuesta se busca modernizar ese acceso, garantizando que la información siga siendo manejada de manera responsable y segura. La idea es contar con una herramienta que organice la información y la convierta en un recurso útil para los funcionarios, facilitando el trabajo diario y permitiendo una administración más ágil y transparente.

Objetivo general: Diseñar e implementar un asistente que permita centralizar, buscar y presentar información institucional de manera rápida y confiable, reduciendo los tiempos de respuesta y apoyando los procesos de gestión del conocimiento dentro del Ministerio.

Objetivos Específicos:

- Preparar un entorno interno y seguro que permita el uso de la herramienta dentro del Ministerio sin depender de servicios externos, cuidando siempre la confidencialidad de la información.
- Asegurar que el asistente consulte únicamente documentos oficiales, para que las respuestas sean confiables, claras y basadas en información vigente.

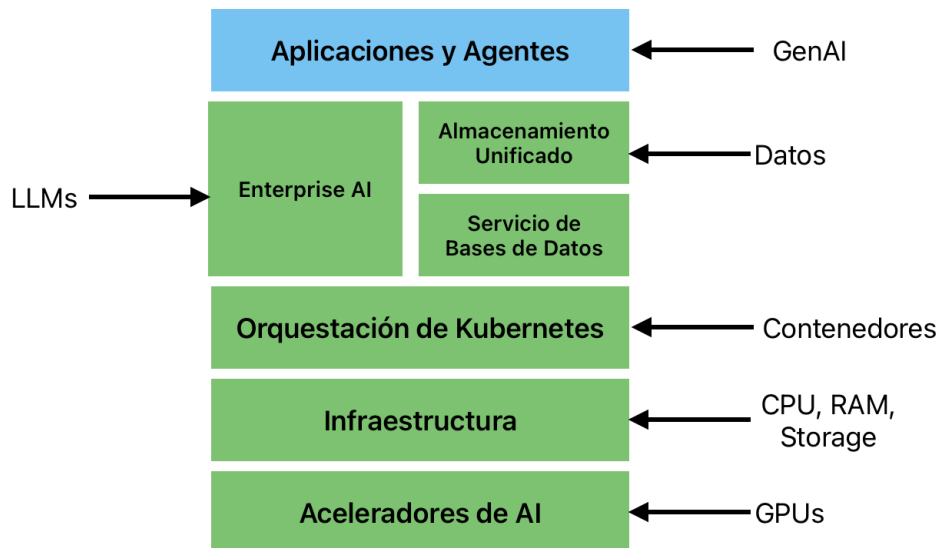
■ Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

- Definir reglas claras para el manejo y protección de los datos, siguiendo la normativa colombiana y buenas prácticas de seguridad, con el fin de garantizar un uso responsable y supervisado de la herramienta.
- Acompañar la adopción del asistente mediante capacitaciones y espacios de retroalimentación, de manera que su uso realmente contribuya a mejorar los procesos del Ministerio y tenga un impacto sostenible.

5.2 Descripción de la solución

La solución propuesta consiste en el despliegue de un ecosistema de IA generativa soberano. A diferencia de las soluciones en la nube, esta arquitectura se ejecuta en servidores propios de la entidad. El núcleo del sistema es un modelo de lenguaje de última generación (como Llama 3 o Mistral) optimizado mediante la técnica RAG (Gao et al., 2024).

El proceso técnico funciona de la siguiente manera: los documentos institucionales (normas, decretos, manuales) son procesados y convertidos en vectores numéricos almacenados en una base de datos vectorial local. Cuando un funcionario realiza una consulta, el sistema busca los fragmentos de información más relevantes y los entrega al LLM para que este genere una respuesta coherente, veraz y con fuentes citadas. Según Huang et al. (2025), este enfoque "On-Premises" es el camino intermedio ideal para preservar la confidencialidad sin sacrificar las capacidades avanzadas de la IA.



Fuente: Elaboración propia

- **Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia**

5.3 Secuencia de actividades para ejecución de la solución

La ejecución de esta propuesta se ha estructurado siguiendo un mapa de ruta lógico que prioriza la estabilidad de la infraestructura y la integridad de los datos antes del despliegue masivo. Este enfoque secuencial permite realizar ajustes iterativos basados en los resultados de cada fase, asegurando que la integración del asistente de IA no interrumpa los procesos críticos de la entidad. La metodología adoptada garantiza que cada paso, desde la adecuación técnica hasta la capacitación del personal, contribuya a la creación de un sistema robusto, escalable y, sobre todo, confiable para el manejo de la memoria institucional del sector público.

Actividad 1. Selección y preparación del equipo de cómputo: Identificación de las capacidades de procesamiento necesarias para que la inteligencia artificial funcione con rapidez dentro de la entidad, procediendo con la instalación de los servidores físicos en el lugar donde residirá el sistema.

Actividad 2. Creación del entorno digital privado: Configuración de un espacio seguro y exclusivo dentro de los sistemas de la organización, asegurando que la herramienta opere de forma aislada y protegida, sin depender de conexiones externas o internet.

Actividad 3. Organización y limpieza de la documentación oficial: Revisión y ordenamiento de los archivos, decretos, manuales, etc. de la entidad para asegurar que la información que se le entregue a la inteligencia artificial sea actual, veraz y libre de errores.

Actividad 4. Estructuración inteligente del conocimiento institucional: Proceso de traducción de los documentos de texto a un formato que el sistema pueda "entender" y consultar de manera casi instantánea, permitiendo que la búsqueda de información sea semántica y no solo por palabras clave.

Actividad 5. Conexión del asistente con la base de datos de la entidad: Configuración de la lógica que permite al asistente de inteligencia artificial leer los documentos internos antes de dar una respuesta, garantizando que sus explicaciones siempre estén fundamentadas en las normas de la organización.

Actividad 6. Diseño del portal de consulta para el funcionario: Creación de una página web sencilla y amigable donde los empleados públicos puedan escribir sus dudas y recibir respuestas del asistente de manera clara y natural.

Actividad 7. Control de calidad y blindaje de la privacidad: Realización de pruebas exhaustivas para verificar que el sistema respete la ley de protección de datos y que las respuestas sean seguras, éticas y alineadas con los valores institucionales.

Actividad 8. Puesta en marcha y entrenamiento de los usuarios: Lanzamiento inicial de la herramienta en un área específica de la entidad y realización de sesiones de acompañamiento para enseñar a los funcionarios a sacar el mayor provecho de esta nueva tecnología.

- Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

5.4 Descripción de recursos necesarios para la ejecución de la solución

Para asegurar la viabilidad y sostenibilidad de la implementación en el contexto de una entidad pública, es fundamental realizar una asignación de recursos multidimensional. No se trata únicamente de un desafío de cómputo, sino de una alineación entre el talento humano, el soporte logístico y el marco financiero que rige la inversión estatal. La correcta identificación de estos elementos permite prever cuellos de botella operativos y garantiza que la solución cuente con el respaldo necesario para operar de manera ininterrumpida, cumpliendo con los estándares de calidad y seguridad exigidos por el Gobierno Digital en Colombia.

Recursos Humanos:

Para garantizar un despliegue exitoso y sostenible, se requiere un equipo interdisciplinario compuesto por:

- Ingeniero de Datos / DevOps, encargado de la integración técnica, automatización, monitoreo y soporte operativo del sistema.
- Arquitecto de IA (Líder técnico), responsable del modelo LLM, la arquitectura RAG, la optimización del rendimiento y el aseguramiento de la calidad en las respuestas del asistente.
- Gestor de Proyecto con experiencia en sector público, encargado de coordinar las actividades, asegurar cumplimiento normativo y facilitar la comunicación entre equipos técnicos y misionales.
- Equipo de apoyo funcional (opcional) para validación de contenido, pruebas y retroalimentación durante el piloto.

Tecnológicos

La solución requiere una infraestructura tecnológica avanzada con capacidad para ejecutar modelos de lenguaje y sistemas de búsqueda semántica de manera segura:

- Estaciones de trabajo de alto rendimiento para el equipo de desarrollo.
- Tablets y dispositivos de consulta para los puntos de atención al funcionario.
- Servidores con capacidad de cómputo gráfico (GPUs de alta gama).
- Software de código abierto para la orquestación de LLMs (LangChain o LlamaIndex).
- Bases de datos vectoriales (ChromaDB o similares).

Infraestructura

Dado que el modelo se ejecutará en los servidores propios del Ministerio, la infraestructura física deberá incluir:

- Espacio físico en el centro de datos de la entidad con las condiciones térmicas y eléctricas adecuadas para el funcionamiento de los servidores.

■ **Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia**

- Centro de datos con certificación de seguridad física.
- Sistema de refrigeración y control térmico de precisión para garantizar la operatividad de los servidores GPU.
- Sistema de alimentación ininterrumpida (UPS) de alta capacidad.

Capacitación

La adopción de la solución requiere un proceso formativo sistemático que facilite la apropiación del asistente por parte de los funcionarios:

- Plataforma de e-learning institucional para el despliegue de módulos de autoaprendizaje sobre el uso de la IA.
- Material pedagógico multimedia (guías interactivas y videotutoriales) sobre ingeniería de prompts para funcionarios públicos.
- Contratación de instructores certificados en ética de la IA y herramientas de productividad generativa.

Financieros

La estimación financiera debe contemplar los costos directos y logísticos asociados a la implementación:

- Equipo técnico (3 especialistas por 6 meses) \$156.000.000
- Servidor GPU, almacenamiento y equipos de oficina \$135.000.000
- Adecuación de datacenter, refrigeración y UPS \$22.000.000
- Capacitación, Talleres, licencias e-learning y material pedagógico \$15.000.000

Total estimado \$328.000.000 COP

Esta inversión asegura la operación segura del sistema, su escalabilidad y su alineación con las políticas de protección de datos y gestión documental aplicables al sector público.

- Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

6 Definición del alcance de la solución, partes interesadas, cronograma y presupuestos.

En esta sección se define el alcance de la solución propuesta, estableciendo sus objetivos, límites y resultados esperados. Asimismo, se identifican las partes interesadas y los responsables de cada actividad dentro del plan de acción, con el fin de garantizar una adecuada coordinación en la ejecución del proyecto. Posteriormente, se presenta el cronograma de actividades que organiza temporalmente las fases de implementación y los hitos de seguimiento. Finalmente, se describe el presupuesto estimado, detallando los recursos financieros requeridos para el desarrollo de la propuesta y su viabilidad.

6.1 Alcance final de la solución

La propuesta se centra en diseñar y poner en funcionamiento un prototipo que ayude a mejorar el acceso y el uso de la información dentro del Ministerio. La idea es contar con una herramienta que permita hacer preguntas de manera sencilla y obtener respuestas claras basadas en los documentos internos de la entidad.

El proyecto se desarrollará inicialmente como un piloto, lo que permitirá comprobar si la herramienta funciona bien y si se adapta a las necesidades del Ministerio antes de pensar en una implementación más amplia. Para ello, se realizarán actividades como organizar la información existente, preparar la herramienta para que pueda usarla el personal, conectarla con los documentos institucionales y hacer pruebas con algunos usuarios.

Además, con esta intervención se busca demostrar cómo una herramienta de este tipo puede ayudar a mejorar los tiempos de consulta, apoyar la toma de decisiones y fortalecer los procesos relacionados con la gestión del conocimiento dentro del Ministerio.

Es importante señalar que esta propuesta se limita únicamente a la fase piloto, cuyo propósito es verificar el funcionamiento de la herramienta y recoger opiniones de los usuarios para mejorarla. Este enfoque coincide con la metodología de investigación-acción, que promueve probar las soluciones en situaciones reales y ajustar lo necesario a partir de la experiencia (Lewin, 1946).

6.2 Identificación de responsables de las actividades para ejecución

La implementación de la propuesta requiere la participación de diferentes grupos dentro del Ministerio, cada uno con un papel importante para que el proyecto avance y pueda ser validado correctamente.

■ Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

En primer lugar, la alta dirección tiene un rol clave, ya que su apoyo permite impulsar la iniciativa y asegurar los recursos necesarios. Además, su respaldo garantiza que el proyecto esté alineado con los planes institucionales y con los esfuerzos de modernización del Ministerio.

El equipo de tecnología también cumple una función esencial, ya que se encarga de preparar los elementos técnicos y de integrar la herramienta con las plataformas que ya usa la entidad.

Por otro lado, los expertos y funcionarios de las áreas misionales aportan su conocimiento sobre los procesos del Ministerio y ayudan a verificar que la información utilizada sea la adecuada. Su participación es especialmente valiosa durante las pruebas, porque permiten identificar mejoras desde la perspectiva de quienes usarán la herramienta en su trabajo diario.

Finalmente, todos los usuarios del Ministerio se benefician directamente de la solución, ya que podrán acceder a la información de una forma más rápida y sencilla.

La participación activa de estos actores refleja el espíritu colaborativo de la investigación-acción, que reconoce la importancia del conocimiento que tienen los propios funcionarios sobre su entorno y sus procesos.

6.3 Cronograma de ejecución

El proyecto está pensado para desarrollarse en un periodo de cinco a seis meses, tiempo durante el cual se llevarán a cabo las distintas etapas necesarias para diseñar, poner en marcha y evaluar el prototipo.

En la primera fase se realizará la planificación, donde se definirá qué se necesita para desarrollar la herramienta, qué información se utilizará y cómo se organizará su funcionamiento general.

Luego se pasará a la preparación e integración de los datos, lo cual incluye ordenar los documentos institucionales y dejarlos listos para que puedan ser utilizados por la herramienta.

En la tercera fase se configurará e implementará el sistema y se harán las pruebas necesarias para asegurar que funcione correctamente.

Una vez listo el prototipo, se pondrá a prueba con usuarios del Ministerio. El objetivo de esta etapa es evaluar cómo se comporta la herramienta en situaciones reales de consulta de información.

Finalmente, se realizará una fase de evaluación y ajustes, en la que se revisarán los resultados, se identificarán mejoras y se prepararán recomendaciones para una posible ampliación del proyecto en el futuro.

■ Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

Tabla 1. Cronograma de actividades.

Actividad	Mes 1	Mes 2	Mes 3	Mes 4	Mes 5	Mes 6
1. Planificación del proyecto y definición de requerimientos	X					
2. Preparación e integración de los datos	X	X				
3. Configuración e implementación del modelo		X	X			
4. Fase de pruebas piloto con usuarios			X	X		
5. Fase de evaluación y ajuste				X	X	
6. Documentación final del proyecto					X	X

Fuente: Creación Propia

6.4 Presupuesto

El presupuesto del proyecto contempla los recursos necesarios para el desarrollo del prototipo, incluyendo talento humano especializado, infraestructura tecnológica y actividades de validación con usuarios. Los valores presentados se expresan en pesos colombianos (COP), considerando que el proyecto se desarrolla en el contexto de una entidad pública ubicada en Colombia.

Dentro de los principales componentes del presupuesto se consideran los costos asociados al equipo técnico encargado del desarrollo de la solución, la adquisición o adecuación de infraestructura tecnológica necesaria para ejecutar el modelo de inteligencia artificial, y las actividades de capacitación y acompañamiento dirigidas a los usuarios que participarán en el piloto.

Es importante señalar que, al tratarse de una fase de experimentación y validación, el presupuesto se enfoca en garantizar los recursos mínimos necesarios para el desarrollo del prototipo. De esta manera, se busca evaluar el impacto potencial de la solución antes de considerar inversiones mayores asociadas a su implementación a escala institucional.

La estimación de costos se realizará considerando criterios de eficiencia en el uso de los recursos disponibles, así como la posibilidad de aprovechar infraestructura tecnológica ya existente dentro del Ministerio.

Tabla 2. Presupuesto.

■ Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

Rubro	Valor
Humanos	
Científico de datos / Ingeniero de IA	\$72.000.000
Ingeniero de datos	\$60.000.000
Ingeniero DevOps / infraestructura	\$32.000.000
Tecnológicos	
Licencias de software y herramientas	\$12.000.000
Infraestructura	
Infraestructura tecnológica (servidor GPU, almacenamiento)	\$85.000.000
Infraestructura adicional / adecuaciones	\$52.000.000 <i>(parte del entorno tecnológico del sistema)</i>
Capacitación	
Actividades de pruebas piloto	\$9.000.000
Capacitación a usuarios	\$6.000.000
Total (COP)	\$328.000.000

Fuente: Creación Propia

El presupuesto estimado para la implementación de la propuesta asciende a \$328.000.000 pesos colombianos (COP), valor que contempla los recursos humanos especializados, la infraestructura tecnológica necesaria para el procesamiento de modelos de inteligencia artificial, la adecuación del entorno físico de operación y las actividades de capacitación dirigidas a los usuarios del sistema. La ejecución del proyecto se plantea en un periodo aproximado de seis meses, durante el cual se desarrollarán las fases de preparación de datos, desarrollo del modelo, implementación del prototipo y validación mediante pruebas piloto. La distribución del presupuesto prioriza la inversión en talento humano y en infraestructura tecnológica, elementos fundamentales en proyectos de inteligencia artificial. En este sentido, la propuesta se considera técnica y financieramente viable, ya que los recursos estimados permiten el desarrollo de un prototipo funcional que facilite la evaluación del impacto de la solución en el contexto organizacional.

- Propuesta de implementación de un asistente de inteligencia artificial generativa On-Premises para la gestión del conocimiento en entidades públicas en Colombia

7 Referencias bibliográficas

- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv. <https://arxiv.org/abs/2312.10997>
- Huang, H., Li, Y., Jiang, B., Liu, L., Sun, R., Liu, Z., & Liang, S. (2025). On-Premises LLM Deployment Demands a Middle Path: Preserving Privacy Without Sacrificing Model Confidentiality. ICLR Workshop: Building Trust in LLMs and LLM Applications. <https://openreview.net/pdf?id=u61yT9ZkEZ>
- Lewin, K. (1946). Action research and minority problems. Journal of Social Issues, 2(4), 34–46. <https://doi.org/10.1111/j.1540-4560.1946.tb02295.x>
- Kemmis, S., & McTaggart, R. (2005). Participatory action research: Communicative action and the public sphere. En N. K. Denzin & Y. S. Lincoln (Eds.), The SAGE handbook of qualitative research (pp. 559–603). SAGE Publications Ltd. <https://researchoutput.csu.edu.au/en/publications/participatory-action-research-communicative-action-and-the-public/>
- Segovia-García, N. y Said-Hung, E.M. (2021). Factores de satisfacción de los alumnos en e-learning en Colombia. *Revista Mexicana de Investigación Educativa*, 26(89), 595-621. <https://cutt.ly/LbUPghi>
- Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology (NIST). <https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10>