Sistema De Consulta (Chatbot) Basado En RAG Para Usuarios Cajas De Compensación

Lucero Alejandra Mojica Tabares

María Elizabeth Pinzón Nova

Cindy Bustamante Serrato

John Jairo Porras

Universidad EAN

Facultad de Ingeniería

Bogotá 23 de febrero 2025

Tabla de contenido

Abstract

Tecnología RAG al Servicio del Afiliado

Este documento presenta una propuesta de automatización de procesos cuyo propósito será implementar un agente virtual (chatbot) basado en tecnología de Generación Aumentada por Recuperación (RAG por sus siglas en inglés), cuyo diseño permitirá consultar información en bases de datos sobre normativa específica de las cajas de compensación y generar respuestas precisas y adecuadas para el afiliado. El sistema integrará un motor de búsqueda apoyado con Inteligencia Artificial (IA) que permitirá extraer fragmentos relevantes de los documentos encontrados para la consulta del usuario.

El Procesamiento del Lenguaje Natural (NLP por sus siglas en inglés) facilitará la estructuración y enriquecimiento del contexto de la respuesta con la información recuperada para hacerla más coherente, detallada y accesible para el usuario. Este agente será útil para la comunidad y brindará educación sobre los beneficios a los que tienen derecho los usuarios de las cajas de compensación y atención al cliente de una manera más oportuna y clara, garantizando que la información entregada en la respuesta para que no sea interpretada como genérica o errada y se proporcione información actualizada mejorando la experiencia del usuario al dar accesibilidad a los servicios.

Palabras Clave:

- Generación Aumentada por Recuperación (RAG)
- Agente virtual (chatbot)
- Procesamiento del Lenguaje Natural (NLP)
- Inteligencia Artificial (IA)
- Automatización de procesos
- Cajas de compensación
- Atención al cliente
- Experiencia del usuario
- Normativa
- Motor de búsqueda

RAG Technology at your service

This document presents a process automation proposal whose purpose will be to implement a virtual agent (chatbot) based on Recovery Augmented Generation (RAG) technology. This design will allow consulting information about specific regulations of compensation funds in databases and generate precise and appropriate responses for the affiliate. The system will integrate a search engine supported by Artificial Intelligence (AI) that will allow to extract relevant fragments from the documents found for the user's query.

Natural Language Processing (NLP) will facilitate the structuring and enrichment of the response context with the retrieved information to make it more coherent, detailed and accessible for the user. This agent will be useful to the community and will provide education for their affiliates about the benefits of compensation funds have for them and fast and agile customer service, ensuring that the information provided in the response is recent, not generic or erroneous, improving the user experience and providing accessibility to services.

Key Words:

- Recovery Augmented Generation (RAG)
- Virtual agent (chatbot)
- Natural Language Processing (NLP)
- Artificial Intelligence (AI)
- Process automation
- Compensation funds
- Customer service
- User experience
- Regulations
- Search engine

Introducción

En la actualidad, la automatización y la inteligencia artificial (IA) han revolucionado la forma en que las personas acceden a la información. Dentro de estas innovaciones, los chatbots han ganado protagonismo en la optimización del servicio al cliente, agilizando la entrega de respuestas y mejorando la experiencia del usuario en distintos sectores (Følstad & Brandtzæg, 2017). Su implementación ha demostrado ser especialmente efectiva en áreas como el comercio electrónico, la banca y la educación, donde la rapidez y precisión en la información son claves para garantizar la satisfacción del usuario (Følstad & Brandtzæg, 2017).

En el caso de las cajas de compensación familiar en Colombia, uno de los principales desafíos para los afiliados es el acceso a la normativa vigente, ya que la información se encuentra fragmentada en diversas plataformas y documentos, lo que dificulta su consulta y comprensión. Esta normativa, que abarca desde los procesos de afiliación hasta la asignación de subsidios y beneficios, requiere constantes actualizaciones y una presentación clara que facilite la toma de decisiones. Para abordar este problema, se plantea el desarrollo de un chatbot basado en tecnología RAG (Retrieval-Augmented Generation), diseñado para centralizar y simplificar la búsqueda de información normativa.

La tecnología RAG combina la recuperación de datos con la generación automatizada de respuestas, permitiendo una interacción más fluida y precisa con los usuarios (Lewis et al., 2020). Este sistema no solo busca y sintetiza información relevante, sino que también proporciona respuestas adaptadas a las consultas específicas de los afiliados. Un aspecto clave para su éxito es la adaptación de la interfaz conversacional a las particularidades del público en Colombia, considerando factores culturales y niveles de alfabetización digital para garantizar una experiencia intuitiva y accesible (Llugsa & Vaca, 2022).

Para lograrlo, es fundamental diseñar un sistema escalable y flexible que integre fuentes normativas de distintas entidades y períodos, asegurando una cobertura completa de la información necesaria para los afiliados (Følstad & Brandtzæg, 2017). Experiencias previas en otros sectores han demostrado que la adopción de chatbots con IA no solo mejora la eficiencia en la atención al usuario, reduciendo tiempos de espera, sino que también permite a los funcionarios enfocarse en casos más complejos, optimizando la gestión del servicio al cliente (Llugsa & Vaca, 2022).

Otro aspecto fundamental es garantizar la privacidad y seguridad de la información. Dado que las cajas de compensación manejan datos personales y sensibles de sus afiliados, es crucial cumplir con la normativa nacional de protección de datos. Esto implica fortalecer la infraestructura tecnológica y establecer protocolos de seguridad que generen confianza en los usuarios y minimicen riesgos asociados a la gestión de grandes volúmenes de información (Piattini & Velthuis, 2020).

Además, la integración de la tecnología RAG con otras herramientas de automatización podría aumentar la eficiencia y precisión en la atención al afiliado. Conectar el chatbot a sistemas de gestión de recursos humanos o plataformas de servicios en línea permitiría una actualización constante de la normativa y un mejor seguimiento de las solicitudes, creando un ecosistema digital más ágil y transparente, capaz de responder eficazmente a las necesidades de los afiliados (Llugsa & Vaca, 2022).

También es importante analizar las herramientas tecnológicas disponibles para la implementación del chatbot, evaluando su viabilidad y compatibilidad con los requerimientos del sistema. Del mismo modo, resulta esencial definir la metodología utilizada en la creación del prototipo, abarcando la selección de datos normativos, el diseño de la arquitectura y la configuración del modelo de IA. Posteriormente, la evaluación de los resultados permitirá identificar oportunidades de mejora y determinar en qué medida la solución cumple con los estándares de accesibilidad, precisión y eficiencia.

El desarrollo de soluciones basadas en IA representa un paso significativo en la modernización de las cajas de compensación familiar, al mejorar la accesibilidad y comprensión de la normativa vigente. Además, la implementación de chatbots con tecnología RAG no solo reduce la carga administrativa derivada de consultas recurrentes, sino que también optimiza el uso de los recursos humanos. Esta iniciativa tiene el potencial de aumentar la satisfacción de los afiliados, proporcionando respuestas más rápidas, detalladas y contextualizadas, contribuyendo así a la evolución de servicios tecnológicos más eficientes y accesibles para toda la comunidad.

Objetivos

Objetivo General

Implementar un prototipo funcional de un chatbot, basado en una arquitectura de Recuperación Aumentada por Generación (RAG) adaptada de una solución de código abierto, con el fin de validar su capacidad técnica para la consulta y comprensión de la normativa aplicable a las Cajas de Compensación Familiar.

Objetivos Específicos

- Desplegar una solución de código abierto (Microsoft RAG OSS) como base arquitectónica del chatbot, utilizando Azure para su establecimiento en un entorno de pruebas.
- Estructurar un corpus documental inicial con normativa específica de las Cajas de Compensación Familiar en el sistema RAG.
- Realizar una evaluación del prototipo del chatbot mediante la ejecución de consultas sobre la normativa cargada, analizando la precisión de las respuestas generadas.

Definición del problema

En el escenario actual de las cajas de compensación familiar en Colombia, el acceso a la normativa regulatoria se ha convertido en un desafío para los afiliados. Estos usuarios suelen enfrentar la dispersión de documentos en múltiples plataformas, la complejidad del lenguaje legal y la falta de medios eficientes para localizar información actualizada (Piattini & Velthuis, 2020). Como consecuencia, la búsqueda y comprensión de beneficios, requisitos y procedimientos normativos resulta poco accesible, generando incertidumbre y una menor utilización de los servicios disponibles.

A pesar de la existencia de portales web y documentos oficiales, la información normativa se encuentra fragmentada y a menudo expuesta en un lenguaje técnico que dificulta su interpretación. Esta situación genera barreras para que los afiliados puedan conocer y ejercer sus derechos de manera efectiva. Asimismo, las cajas de compensación están obligadas a mantener actualizada la normativa y responder adecuadamente a las consultas de sus afiliados; sin embargo, la falta de mecanismos eficaces de acceso a la información conlleva a un aumento en la carga administrativa y a la insatisfacción de los usuarios (Følstad & Brandtzæg, 2017).

En este contexto, surgen varios cuestionamientos que delimitan el problema de investigación:

- ¿Cuáles son las principales dificultades que enfrentan los afiliados para acceder y comprender la información normativa de las cajas de compensación familiar?
- ¿Qué factores contribuyen a la dispersión y fragmentación de la información normativa en los portales y documentos oficiales?
- ¿Cómo afecta la falta de acceso eficiente a la información normativa en la toma de decisiones y uso de los servicios por parte de los afiliados?
- ¿Cuál es el impacto de la ausencia de un sistema unificado de consulta sobre la carga administrativa de las cajas de compensación?

Para responder a estas preguntas, la investigación se basará en la recolección y análisis de datos empíricos, como registros de consultas, encuestas a usuarios y evaluaciones de accesibilidad de la información normativa. Se aplicarán criterios de confiabilidad en la recolección de datos, asegurando que las métricas utilizadas permitan evaluar de manera objetiva las dificultades de acceso a la información normativa. Además, se considerarán los

aspectos éticos en la recopilación y manejo de datos de los usuarios, garantizando la privacidad y el consentimiento informado.

El ámbito temporal del estudio se enfocará en la normativa vigente entre 2020 y 2025, permitiendo analizar la evolución reciente de la legislación aplicable a las cajas de compensación familiar. Geográficamente, el estudio se llevará a cabo en el contexto colombiano, con énfasis en la ciudad de Bogotá, donde se concentra un alto número de afiliados formales (Mercado laboral de Bogotá, 2023). La población de estudio estará compuesta por personas afiliadas de entre 18 y 60 años, quienes regularmente consultan beneficios o procedimientos legales en estas entidades.

El impacto potencial de esta investigación se extiende más allá del contexto inmediato, ya que sus hallazgos podrán servir de base para futuras estrategias de digitalización de información regulatoria en Colombia. Asimismo, podría contribuir al diseño de herramientas que mejoren la experiencia de usuario y reduzcan la carga administrativa en otras entidades con problemas similares de acceso a la información. De esta manera, la investigación se orienta a una problemática concreta y busca generar conocimiento sobre las barreras existentes en el acceso a la información normativa, sin adelantar soluciones específicas, pero proporcionando una base sólida para el desarrollo de estrategias futuras que permitan mejorar la situación.

Justificación

En la actualidad, las cajas de compensación familiar enfrentan un reto creciente al momento de difundir y mantener actualizada la normativa que rige sus servicios. La información suele encontrarse dispersa en documentos o portales diferentes, lo que dificulta la comprensión por parte de los afiliados y genera confusión al momento de solicitar sus beneficios (Piattini & Velthuis, 2020). Frente a esta situación, un chatbot basado en Recuperación Aumentada por Generación (RAG) adquiere relevancia porque centraliza de manera inteligente los datos legales y responde a las consultas de los usuarios de forma inmediata y contextualizada (Lewis et al., 2020).

Desde la perspectiva de los afiliados, contar con una herramienta de IA que interprete sus preguntas en lenguaje natural y ofrezca respuestas precisas les brinda autonomía y reduce la necesidad de recurrir a canales convencionales de atención, que suelen saturarse con preguntas repetitivas (Følstad & Brandtzæg, 2017). Para las cajas de compensación, este enfoque implica liberar recursos humanos, concentrándolos en casos o solicitudes que requieran un análisis más detallado y personalizado. Además, propicia una mejor experiencia de usuario, ya que los afiliados podrán acceder a la información normativa sin tener que navegar por múltiples fuentes o interpretar términos legales complejos (Llugsa & Vaca, 2022).

Al abordar esta problemática con tecnología RAG, el proyecto no solo aspira a agilizar y eficientar la búsqueda de información normativa, sino también a sentar las bases de una transformación digital más profunda. El impacto positivo se extiende, por un lado, al empoderamiento ciudadano, pues las personas tendrán un mayor control y conocimiento de los beneficios a los que pueden acceder; por otro lado, las instituciones verán fortalecida su capacidad de respuesta y su imagen de modernidad e innovación. Así, la implementación de esta solución abre la posibilidad de replicarse en otras entidades con retos normativos similares, ampliando su alcance y consolidando la adopción de la IA en el ámbito legal (Lewis et al., 2020).

Finalmente, esta iniciativa tiene un claro componente de investigación, ya que integra técnicas de recuperación de datos y generación de texto en un contexto muy específico y crítico: la normativa legal de las cajas de compensación. A través de la evaluación empírica del chatbot y la recopilación de datos sobre su uso, se logrará no solo validar la eficacia de la

tecnología RAG en el entorno colombiano, sino también contribuir al desarrollo de mejores prácticas de diseño y entrenamiento de modelos de lenguaje natural. Así, el proyecto pretende aportar tanto al bienestar de la comunidad afiliada como al acervo de conocimiento tecnológico y académico en el país.

Análisis de Requerimientos

1. Requerimientos Funcionales

- Uso de Lenguaje Natural
  - Entender y procesar preguntas formuladas por los usuarios en lenguaje natural sobre las cajas de compensación.
  - Estructurar y enriquecer el contexto de las respuestas.
  - Interpretación de consultas complejas y ambigüas.

- Precisión en las respuestas
  - Proporcionar respuestas precisas y claras, evitando la interpretación de términos legales complejos basadas en la información normativa consultada.
  - Generar respuestas coherentes, detalladas y accesibles para el usuario.
  - Evitar generar respuestas genéricas o erróneas.
  - Se necesita asegurar que las respuestas entregadas contengan información actualizada.

- Consulta y Recuperación de Información Normativa:
  - Integrar y centralizar la normativa dispersa, ofreciendo respuestas rápidas y coherentes usando un motor de búsqueda apoyado con IA para extraer fragmentos relevantes de los documentos consultados en múltiples bases de datos y formatos.
  - Los usuarios deben poder acceder a la información normativa sin necesidad de navegar por múltiples fuentes.
  - El sistema debe permitir a los usuarios consultar información sobre normativa específica de las cajas de compensación.
  - Priorizar la información más relevante para la consulta del usuario.

- Interfaz de Usuario:
  - Se debe tener una interfaz de usuario intuitiva y fácil de usar.
  - Accesibilidad desde múltiples dispositivos (computadoras, teléfonos móviles, etc.).
  - Proporcionar una experiencia de usuario fluida y satisfactoria.

2. Requerimientos No Funcionales:
   - Rendimiento:
     - El chatbot debe responder a las consultas de los usuarios de manera rápida y eficiente.
     - El motor de búsqueda debe ser capaz de procesar grandes volúmenes de datos de manera eficiente.
     - El sistema debe ser escalable para manejar un gran volumen de consultas y respuestas.
     - El chatbot debe ser rápido en la recuperación de información y generación de respuestas.
     - Debe poder manejar múltiples consultas simultáneas sin afectar su rendimiento.

   - Seguridad
     - El sistema debe proteger la información confidencial de los usuarios y las cajas de compensación.
     - Debe cumplir con las normas de seguridad y privacidad de datos aplicables.
     - Es necesario asegurar que la información del usuario no se almacene ni se utilice de forma indebida.

   - Documentación:
     - El código fuente debe estar bien documentado y organizado.

3. Requerimientos Técnicos

   - Utilización de una solución de código abierto para la recuperación de información desde un corpus documental normativo y la generación de respuestas contextualizadas.

   - Soporte para consultas en lenguaje natural con un enfoque conversacional.

   - Desarrollar en una plataforma que permita la integración de tecnologías de IA y NLP.

Marco de referencia

Tema y Subtema	Teoría/Modelo/Concepto	Descripción o Idea Central	Autor y Año	Fuente APA
1. Gestión del Conocimiento en Organizaciones	Modelo SECI	Proceso de creación y transferencia de conocimiento en organizaciones	Nonaka & Takeuchi (1995)	Nonaka, I., & Takeuchi, H. (1995). *The Knowledge-Creating Company*. Oxford University Press.
1.1. Sistemas de Gestión del Conocimiento	Teoría de Gestión del Conocimiento	Implementación de sistemas para almacenamiento y acceso eficiente a información	Alavi & Leidner (2001)	Alavi, M., & Leidner, D. E. (2001). Knowledge management systems: Foundations. MIS Quarterly, 25(1), 107-136.
1.2. Desafíos en entornos dinámicos	Limitaciones de SGC tradicionales	Retos en gestión de conocimiento en entornos normativos cambiantes	Hislop (2013)	Hislop, D. (2013). Knowledge management in organizations. Oxford University Press.
2. Tecnología RAG	Generación Aumentada por Recuperación	Combina recuperación de información con generación de respuestas	Lewis et al. (2020)	Lewis, P., et al. (2020). Retrieval-augmented generation. NeurIPS, 33, 9459-9474.
2.1. Aplicación corporativa de RAG	Implementación en entornos normativos	Solución para consultas precisas en documentación fragmentada	Borgefalk et al. (2023)	Borgefalk, P., et al. (2023). Legal information retrieval using RAG. JAIR, 45(2), 123-145.
3. Chatbots y PLN	Interacción Humano-Computadora	Sistemas conversacionales para automatización de servicios	Følstad & Brandtzaeg (2017)	Følstad, A., & Brandtzaeg, P. B. (2017). Chatbots in HCI. Interactions, 24(4), 38-42.

Tema y Subtema	Teoría/Modelo/Concepto	Descripción o Idea Central	Autor y Año	Fuente APA
3.1. Impacto organizacional	Eficiencia operativa	Reducción de tiempos en consultas recurrentes	Abdellatif et al. (2022)	Abdellatif, A., et al. (2022). Chatbot implementation challenges. Journal of Business Tech, 17(3), 45-67.
4. Seguridad y Privacidad	Normativas de protección de datos	Cumplimiento de estándares en sistemas de IA	Yang et al. (2023)	Yang, J., et al. (2023). Security in chatbots. Applied Sciences, 13(11), 6355.
4.1. Aspectos éticos en IA	Principios de IA confiable	Guías para implementación ética de sistemas de IA	Floridi et al. (2021)	Floridi, L., et al. (2021). Ethical AI guidelines. AI Ethics Journal, 3(1), 56-78.

La gestión del conocimiento en organizaciones complejas como las cajas de compensación familiar enfrenta desafíos únicos debido a la naturaleza dinámica y fragmentada de la información normativa. El modelo SECI (Nonaka & Takeuchi, 1995) ofrece un marco conceptual robusto para entender cómo el conocimiento se crea, comparte y utiliza en las organizaciones. Este modelo destaca cuatro procesos clave: socialización (compartir conocimiento tácito), externalización (convertir conocimiento tácito en explícito), combinación (integrar conocimiento explícito) e internalización (asimilar conocimiento explícito). En el contexto de las cajas de compensación, este marco ayuda a entender cómo el conocimiento sobre normativas y procedimientos puede fluir eficientemente entre diferentes áreas y niveles organizacionales.

Los Sistemas de Gestión del Conocimiento (SGC) tradicionales (Alavi & Leidner, 2001) han demostrado limitaciones significativas en entornos donde la información cambia constantemente. Mientras que estos sistemas son efectivos para almacenar conocimiento

explícito, a menudo fallan en capturar conocimiento tácito y en proporcionar mecanismos ágiles para actualizar y recuperar información. Esto es particularmente problemático en el sector de compensación familiar, donde las normativas pueden cambiar frecuentemente y donde diferentes usuarios (empleados, afiliados, administradores) necesitan acceso a versiones actualizadas de la información (Hislop, 2013). La implementación efectiva de SGC en este contexto requiere no solo soluciones tecnológicas adecuadas, sino también cambios organizacionales y culturales que fomenten el intercambio continuo de conocimiento.

La tecnología de Generación Aumentada por Recuperación (RAG) representa un avance significativo en el campo de la inteligencia artificial aplicada a la gestión de conocimiento (Lewis et al., 2020). A diferencia de los modelos de lenguaje tradicionales que dependen únicamente de su conocimiento pre-entrenado, los sistemas RAG combinan capacidades de recuperación de información (como motores de búsqueda) con capacidades de generación de lenguaje natural. Esto permite que el sistema acceda a bases de conocimiento actualizadas en tiempo real y genere respuestas precisas y contextualizadas. En el contexto de las cajas de compensación, esto significa que un chatbot basado en RAG puede proporcionar información exacta sobre normativas recién publicadas sin necesidad de reentrenamiento constante del model.

La aplicación de RAG en entornos corporativos (Pinzón et al., 2024) ofrece ventajas significativas sobre soluciones tradicionales. Primero, reduce el tiempo y costo asociado con la actualización de bases de conocimiento, ya que el sistema puede acceder directamente a documentos fuente actualizados. Segundo, mejora la precisión de las respuestas al basarse en información verificable en lugar de dependender exclusivamente del conocimiento pre-entrenado del modelo. Tercero, permite mayor transparencia, ya que el sistema puede proporcionar referencias a los documentos fuente utilizados para generar cada respuesta.

Estos beneficios son particularmente valiosos en el contexto normativo de las cajas de compensación, donde la exactitud y actualidad de la información son críticas (Borgefalk et al., 2023).

Los sistemas conversacionales basados en PLN han evolucionado significativamente en los últimos años, pasando de simples árboles de decisión a complejos modelos de inteligencia artificial capaces de entender contexto y matices del lenguaje humano (Følstad & Brandtzaeg, 2017). En el contexto organizacional, los chatbots modernos pueden manejar consultas complejas, mantener el contexto a través de múltiples interacciones y adaptar sus respuestas al nivel de conocimiento del usuario. Para las cajas de compensación, esto significa poder ofrecer un servicio personalizado a diversos usuarios, desde empleados que necesitan información técnica precisa hasta afiliados que requieren explicaciones simples sobre sus beneficios.

El impacto de estos sistemas en la eficiencia organizacional es significativo. Estudios muestran que los chatbots pueden manejar hasta el 80% de las consultas rutinarias, liberando al personal humano para tareas más complejas que requieren juicio y experiencia. Además, al proporcionar respuestas consistentes basadas en documentación oficial, reducen los errores causados por interpretaciones personales de normativas complejas. Sin embargo, el éxito de estas implementaciones depende críticamente de la calidad del diseño de la experiencia de usuario y de la integración con los sistemas y procesos organizacionales existentes (Abdellatif et al., 2022).

La implementación de sistemas de IA que manejan datos sensibles en el sector de compensación familiar requiere un enfoque riguroso en seguridad y privacidad (Yang et al., 2023). La norma ISO 27001 proporciona un marco completo para establecer, implementar, mantener y mejorar continuamente un sistema de gestión de seguridad de la información. En

el contexto colombiano, el Decreto 1377 de 2013 establece requisitos específicos para la protección de datos personales, incluyendo principios de finalidad, libertad, veracidad, acceso y circulación restringida, seguridad y confidencialidad. Estos marcos son esenciales para garantizar que los chatbots basados en IA manejen adecuadamente la información sensible de los afiliados.

Además de los requisitos normativos, es fundamental considerar aspectos éticos en el diseño e implementación de estos sistemas (Floridi et al., 2021). Esto incluye garantizar la transparencia en el funcionamiento del sistema, establecer mecanismos claros de rendición de cuentas, y prevenir sesgos en las respuestas generadas. Las cajas de compensación, como entidades de carácter social, tienen una responsabilidad adicional de garantizar que estas tecnologías sean accesibles e inclusivas para todos sus afiliados, independientemente de su nivel de alfabetización digital.

Análisis de Restricciones

Restricciones técnicas

- Procesamiento de datos: limitación de costos por uso de infraestructura en nube o de recursos por infraestructura propia para el procesamiento de altos volúmenes de conectividad, datos y respuestas hacia el afiliado.
- Capacidad de Adaptación: complejidad en tiempos y costos de actualización y entrenamiento de los modelos de IA que pueden degradar la calidad de la interacción y las respuestas
- Seguridad de la Información: Inadecuada implementación/actualización de herramientas de seguridad, control no efectivo de accesos, no aplicación de normativa y procesos de protección de datos.

- Actualización de información: Bases de datos con información no especializada, desactualizada, incompleta, ilegible o ambigua.

Restricciones de usabilidad

- Uso de multiplataforma: Limitación del acceso a la herramienta desde diversos dispositivos como, PC, teléfonos móviles, tabletas, etc.
- Interfaz de usuario: Degradación de la experiencia de usuario por ser poco amigable, difícil de entender o demasiado básica para su interacción con el afiliado.
- Precisión en las respuestas: Fallas o escogencia inadecuada en las tecnologías de infraestructura, IA y/o NLP seleccionadas que puedan presentar imprecisión o demoras en las respuestas hacia el afiliado.

Restricciones de privacidad

- Confidencialidad del usuario: Almacenamiento no autorizado de la información suministrada por el afiliado en la interacción con la herramienta.
- Cumplimiento normativo: No cumplimiento de la normativa vigente o falta de adaptabilidad en la configuración hacia la normativa futura sobre uso de IA o protección de datos.

Restricciones Financieras

- Licencias y herramientas: Costos asociados a licenciamiento de modelos de IA, propiedad intelectual de bases de datos especializadas, plataformas de NLP o RAG.
- Mantenimiento y actualización: Costos recurrentes para el mantenimiento del sistema y la actualización de la normativa o la infraestructura.

Restricciones de Tiempo

- Tiempo de implmentación: Complejidad de despliegue o falta de uso de metodologías ágiles en el proceso.
- Capacitación del modelo: Periodos demasiado extensos para el entrenamiento del módulo de IA de la herramienta.

Metodología para la selección y desarrollo de la solución

La presente propuesta detalla la metodología para identificar, evaluar y seleccionar la solución tecnológica más adecuada para un chatbot con capacidad de Recuperación Aumentada por Generación (RAG), destinado a las Cajas de Compensación. El proceso se adhiere a principios de ingeniería que priorizan la viabilidad, la eficiencia y el impacto integral (económico, social y ambiental).

Fases Metodológicas

Definición del Problema y Restricciones

Se parte de la necesidad de un chatbot que provea información normativa precisa y actualizada. Las restricciones incluyen la necesidad de coherencia normativa, cumplimiento ético-legal, una óptima experiencia de usuario, y la capacidad de automatización administrativa para Colsubsidio, Compensar y Comfama.

Generación y Selección Preliminar de Alternativas

Verificación de Coherencia Lógica.

Se descartan de antemano enfoques conceptualmente ilógicos o que contravengan principios establecidos (ej. un LLM sin mecanismo de RAG para esta tarea se considera inviable por su propensión a "alucinaciones" y falta de anclaje en fuentes normativas actualizadas, lo cual es una forma de "solución ilógica" para el problema planteado).

Identificación de Alternativas Viables.

Con base en la investigación de tecnologías actuales y "hechos conocidos" (experiencias previas y soluciones implementadas en el sector o con tecnologías similares), se han identificado tres arquitecturas principales como candidatas viables. Esto cumple con la recomendación de contar con un grupo inicial de al menos tres posibilidades para un análisis comparativo robusto.

Alternativa 1.

Solución Cloud Integrada (Proveedor Único, ej. Azure/AWS).

Alternativa 2.

Solución Open-Source Autogestionada (Cloud híbrida o On-Premise).

Alternativa 3.

Plataforma Híbrida con IA Especializada (ej. IBM Watsonx, Google Dialogflow CX).

Calidad de la Información.

La investigación y fundamentación de cada alternativa se basa en referencias académicas, documentación de proveedores tecnológicos (referencias comerciales especializadas) y análisis de casos de uso similares, verificando la especialización de las fuentes.

Evaluación Detallada de Alternativas

Cada una de las tres alternativas preseleccionadas se evalúa rigurosamente frente a un conjunto de criterios clave, derivados de los requerimientos y la "función objetivo" del proyecto:

Coherencia Normativa.

Capacidad de la solución para proveer información alineada con los documentos fuente y facilidad de actualización.

Cumplimiento Ético y Legal.

Adhesión a normativas de protección de datos, transparencia, explicabilidad y mitigación de sesgos.

Experiencia de Usuario (UX) y Automatización Administrativa.

Facilidad de uso, disponibilidad multicanal, tiempos de respuesta y potencial de integración para automatizar procesos.

Costos y Recursos Necesarios.

Inversión inicial, costos operativos, dependencia de proveedores y requerimientos de personal técnico.

Escalabilidad y Mantenibilidad.

Capacidad de crecimiento y facilidad para realizar ajustes y actualizaciones.

Este análisis busca identificar rápidamente las fortalezas y debilidades de cada opción, permitiendo un descarte temprano de las menos favorables para evitar incurrir en gastos innecesarios de tiempo y recursos en ellas.

Análisis Comparativo y Selección de la Solución Recomendada

Se realizará una comparación directa de las alternativas utilizando una matriz de evaluación que resume su desempeño en cada criterio. La selección final se basará en la alternativa que ofrezca el mejor equilibrio, apuntando a la solución más rentable en términos económicos, y con el mayor impacto positivo social y ambiental. Aunque una solución pueda ser descartada inicialmente, se contempla la posibilidad de reconsiderarla si variaciones o nueva información la vuelven competitiva.

Definición de Indicadores de Éxito.

Para medir la efectividad de la solución implementada, se establecerán indicadores clave (KPIs) en las dimensiones de

Rentabilidad.

ROI, reducción de costos operativos, ahorro de tiempo.

Impacto Social.

Satisfacción del usuario, NPS, accesibilidad, empoderamiento ciudadano.

Impacto Medioambiental.

Reducción de huella de carbono por virtualización de trámites, uso de infraestructura eficiente. El indicador principal del proyecto será la proporción de consultas normativas atendidas satisfactoriamente de forma automatizada.

Aplicación de la Metodología: Análisis de Alternativas

A continuación, se presenta un resumen del análisis realizado para las tres alternativas identificadas, siguiendo la metodología expuesta:

Alternativa 1: Solución Cloud Integrada (Proveedor único - Azure/AWS).

Descripción.

Utilización de servicios PaaS/SaaS de un gran proveedor cloud (ej. Azure Cognitive Search, Azure OpenAI, Azure Bot Service) para todo el flujo RAG.

Evaluación Resumida.

Coherencia Normativa.

Muy alta, gracias a motores de búsqueda robustos y LLMs potentes. Fácil actualización del corpus.

Cumplimiento Ético-Legal.

Robusto, apalancado en certificaciones y políticas del proveedor. Requiere gestión de la dependencia y configuración adecuada de privacidad.

UX y Automatización.

Moderna, multicanal, escalable. Integración facilitada con otros servicios cloud y APIs.

Alternativa 2: Solución Open-Source Autogestionada (Cloud híbrida o On-Premise).

Descripción.

Construcción con componentes open-source desplegados en infraestructura propia o cloud controlada.

Evaluación Resumida.

Coherencia Normativa.

Potencialmente alta con personalización profunda y fine-tuning. Mayor control, pero requiere más esfuerzo de ajuste.

Cumplimiento Ético-Legal.

Máxima soberanía de datos. Responsabilidad directa de la organización en implementar gobernanza de IA.

UX y Automatización.

Altamente personalizable. Se provee una interfaz gráfica por defecto. Escalabilidad gestionada internamente.

Alternativa 3: Plataforma Híbrida con IA Especializada (Ej: IBM Watsonx, Google Dialogflow CX).

Descripción.

Uso de plataformas de chatbot empresarial para el front-end y gestión de diálogo, integrando capacidades RAG a medida o mediante conectores.

Evaluación Resumida.

Coherencia Normativa.

Muy alta para FAQs y alta con RAG bien integrado. Plataformas suelen garantizar contexto actualizado.

Cumplimiento Ético-Legal.

Robusto, con funciones de gobierno de IA del proveedor. Requiere contrato y gestión de la relación con el proveedor.

UX y Automatización.

UX refinada de inicio, multicanal nativo. Integraciones pre-hechas pueden acelerar la automatización.

Comparativa de las alternativas

Para visualizar las diferencias clave, la siguiente tabla resume cómo se posiciona cada alternativa en los criterios principales:

Criterio	**Alt 1: Cloud Integrada (Azure/AWS)**	**Alt 2: Open-Source Autogestionada**	**Alt 3: Plataforma Híbrida (IBM/Google)**

Coherencia Normativa	Muy alta (buscadores empresariales + GPT-4 garantizan respuestas precisas con citas) . Fácil de actualizar con nuevas normas.	Alta, con control total (embeddings ajustados al dominio, posibilidad de fine-tuning). Permite mayor personalización.	Muy alta en FAQs conocidas (curadas manualmente). En consultas nuevas, alta gracias a RAG y supervisión de plataforma. Plataformas garantizan contexto actualizado 24/7
Ética y Legal	Cumplimiento robusto vía proveedor (seguridad certificada, acuerdos de datos). Dependencia del proveedor en políticas de IA. Transparencia configurable (citas, disclaimers).	Cumplimiento gestionado internamente: máxima soberanía de datos. Exige gobierno de IA propio (mitigar sesgos, monitorear). Sin terceros con acceso a información.	Cumplimiento robusto vía proveedor (IA responsable incorporada, auditabilidad). Requiere contrato comercial. Plataforma guía mejores prácticas éticas (p.ej. registro, privacidad)
UX y Automatización	UX moderna, multicanal rápido. Integración nativa con servicios cloud y API internas. Escalable automáticamente. Personalización media (limitada a opciones del proveedor).	Provee una interfaz gráfica ya definida, intuitiva y simple.	UX muy refinada de inicio (plantillas, voz). Multicanal nativo. Integraciones pre-hechas aceleran automatización. Permite multi-institución fácilmente. Menor esfuerzo de desarrollo en frontend, más en la configuración.

Costo y Recursos	Modelo SaaS/consumo: costos operativos según uso (puede ser elevado con muchas consultas, pero sin inversión inicial fuerte). Ahorra en mantenimiento de infraestructura.	Requiere inversión en infraestructura (servidores o instancias cloud) y equipo técnico para mantenimiento. Costos fijos pueden ser altos inicialmente, pero bajos variables. Sin licencias propietarias (coste cero de software).	Costos de licencia o suscripción a la plataforma. Soporte incluido. Desarrollo más rápido (menores costos de desarrollo custom). Balance entre costo fijo de plataforma y ahorro en tiempo de implementación.
Indicadores de éxito	Ahorro de tiempo para usuarios y empleados (medible en reducción de consultas manuales) . Escalabilidad (número de consultas atendidas sin degradación). ROI positivo si reduce llamadas tradicionales.	Calidad de respuestas (precisión medida por expertos) comparable a sistemas comerciales. Bajo porcentaje de fallas críticas. ROI en el largo plazo por ahorros de licencias. Impacto social al empoderar con tecnología abierta.	Alta satisfacción del usuario (meta >90% parecido a caso Redbridge). Tiempo de implementación corto (beneficio rápido). Ahorro significativo en atención (costo por consulta reducido, similar a 90% ahorro reportado)

En base al análisis realizado, recomendamos la **Alternativa 2 (Open-Source Autogestionada)** como la opción más adecuada para este proyecto. Esta solución se presenta como una plataforma lista para usar, que requiere un despliegue sencillo y la incorporación de los documentos normativos propios, sin necesidad de un trabajo pesado de desarrollo desde cero. Además, es altamente personalizable; por ejemplo, mediante la clonación del repositorio base y la creación de ramas específicas para adaptar el chatbot a nuestras

necesidades particulares, lo que permite mantener el control total sobre la implementación y futuras mejoras.

Esta alternativa ofrece un balance ideal entre flexibilidad, control y rapidez de puesta en marcha, evitando depender de licencias costosas o plataformas cerradas. Aunque requiere cierto conocimiento técnico para la configuración y mantenimiento, este esfuerzo es moderado y superado ampliamente por la ventaja de poder adaptar el sistema exactamente a los requerimientos de cada Caja, sin ataduras a proveedores externos.

Criterios de descarte de enfoques inviables

- En el proceso de diseño se descartaron además enfoques que, por diversas razones técnicas o funcionales, se consideraron inviables o poco recomendables:
- **LLM sin recuperación (solo generación):** Rechazado por su alta propensión a generar respuestas incorrectas o "alucinaciones" y la incapacidad de actualizarse dinámicamente con cambios normativos. Su falta de contexto actual y acceso a fuentes oficiales lo hace inadecuado para un entorno legal, donde la precisión es crítica (NVIDIA, 2023).
- **Chatbots basados únicamente en reglas o árboles de decisión:** A pesar de su coherencia lógica, son poco escalables y difíciles de mantener en un escenario normativo tan amplio y cambiante. Además, limitan la interacción a preguntas rígidas y respuestas predefinidas, afectando la experiencia del usuario y la capacidad del sistema para manejar consultas complejas.
- **Chatbot único para las tres Cajas:** Debido a las diferencias en reglamentos, beneficios y convenios propios de Colsubsidio, Compensar y Comfama, un solo bot no lograría diferenciar ni contextualizar adecuadamente las respuestas. Por ello, se opta por una base común nacional con instancias personalizadas para cada entidad, garantizando precisión y relevancia.

Indicadores propuestos de rentabilidad, impacto social y medioambiental

Para evaluar el éxito del chatbot RAG una vez implementado, definimos indicadores clave en tres dimensiones:

Rentabilidad (ROI y eficiencia):

- Porcentaje de consultas atendidas por el chatbot vs. canales tradicionales (meta: >80%) (Administració Oberta de Catalunya, 2024)
- Horas-hombre ahorradas y reducción de costos operativos.
- Reducción en tiempo de capacitación de nuevos empleados.
- Cálculo de ROI: ahorros totales comparados con la inversión, con meta de amortización en 1-2 años.

Impacto social:

- Satisfacción del usuario (>90%) mediante encuestas y análisis de sentimiento (IBM, s.f.).
- Net Promoter Score (NPS) como indicador de recomendación del servicio.
- Alcance: número de usuarios únicos y frecuencia de uso.
- Inclusividad: uso en horarios no laborales o por poblaciones con difícil acceso físico.
- Empoderamiento ciudadano: número de trabajadores que aclaran derechos y deberes a través del bot.

Impacto medioambiental:

- Reducción de trámites físicos y desplazamientos, estimando CO_2 evitado por consultas virtuales.
- Uso de infraestructura cloud eficiente y con energía renovable (AWS, Azure) (Sustainability Magazine, 2023).
- Consumo energético rastreado y convertido a métricas de sostenibilidad.
- Minimización de hardware físico nuevo, promoviendo accesos web y reduciendo residuos electrónicos.

Indicador principal del proyecto:

Proporción de consultas normativas atendidas satisfactoriamente de forma automatizada, como medida de adopción y efectividad (Inbenta, 2023). Todos los demás indicadores estarán alineados a esta métrica central.

Arquitectura general de un sistema RAG.

Un modelo de recuperación obtiene información relevante de fuentes internas estructuradas (bases de datos) y no estructuradas (documentos), incorporándola como contexto para que el modelo generativo (LLM) produzca una respuesta precisa y fundamentada. El flujo inicia con la consulta del usuario, pasa por la capa de recuperación que **"inyecta"** datos pertinentes, y finaliza con la generación de la respuesta final (K2view, 2024).

Plan de Implementación y Evaluación

Tras seleccionar la **Alternativa 2 (solución open-source autogestionada)**, el equipo optó por aprovechar un proyecto de código abierto desarrollado por Microsoft, que ofrece una arquitectura RAG preconstruida, lista para desplegar y utilizar. Este enfoque permitió avanzar rápidamente en la implementación sin necesidad de desarrollar desde cero, manteniendo a la vez el control técnico y la capacidad de personalización.

La solución fue adaptada mediante la carga de documentos normativos específicos y configuraciones menores para asegurar su adecuación al contexto colombiano. A continuación, se detalla el plan de implementación ejecutado.

Fase 1: Estrategia de Desarrollo Iterativo y Centrado en el Usuario.

Iteraciones Ágiles de Mejora.

Se adoptará un enfoque de desarrollo ágil. Se mantendrá un backlog dinámico de mejoras (ej. incorporación de nueva normativa, ajustes de redacción, nuevas funcionalidades como descarga de documentos). Estas mejoras se implementarán en ciclos cortos de desarrollo (ej. mensuales), facilitados por la naturaleza cloud-first de la solución base.

Enfoque Adaptable y Retroalimentación Continua.

Se reconocerá que el uso real del sistema revelará necesidades no anticipadas. Por ello, se planifica solicitar y analizar la participación constante de usuarios finales y expertos en la materia para guiar la evolución del chatbot. Se buscará una prevención proactiva de errores estructurales (fuentes incorrectas, contexto insuficiente, etc.) con el objetivo de lanzar una herramienta confiable y con capacidad de mejora continua.

Fase 2: Esquema de Validación y Pruebas Detallado.

Se implementará un esquema de validación y pruebas exhaustivo, tanto en etapas de pre-producción como durante la operación, para asegurar la precisión, eficiencia, claridad y robustez del chatbot RAG.

Pruebas de Recuperación Normativa (Unitarias).

Se definirán queries con respuestas conocidas y se medirá el `recall@K` (ej. `recall@5 > 95%`) para asegurar que los documentos correctos figuren entre los primeros resultados.

Pruebas Funcionales de Generación (End-to-End).

Las respuestas serán evaluadas por expertos legales y usuarios piloto, considerando: exactitud legal (meta: 100%), integridad, claridad (escala 1-5, meta: ≥4/5), tono y formalidad. Se buscará un umbral de precisión global ≥90%.

Pruebas de Rendimiento y Estrés.

Se simulará la concurrencia de múltiples usuarios para medir tiempos de respuesta (meta: <3s normal, <5s con carga) y estabilidad.

Pruebas de Usabilidad (UX).

Se realizarán pruebas con usuarios externos para evaluar la facilidad de uso y se ajustarán elementos de la interfaz según los hallazgos.

Pruebas de Casos Límite y Éticos.

Se evaluará el manejo de ambigüedades, multi-intención, lenguaje inapropiado y consultas fuera de alcance, verificando la alineación con principios éticos y legales.

Fase 3: Adecuación y Carga del Corpus Normativo Específico.

Recopilación y Estructuración del Corpus.

Se procederá a la recopilación exhaustiva de documentos legales y normativos pertinentes a las Cajas de Compensación Colombianas (Código Sustantivo del Trabajo, Ley 100, reglamentos internos, circulares, etc.). Estos documentos se organizarán temáticamente y se les asignarán metadatos relevantes (nombre, fecha, entidad) para optimizar su indexación y recuperación por el sistema.

Ingesta en el Sistema y Validación de Indexación.

Los documentos del corpus se cargarán en el sistema a través del proceso de ingestión definido (basado en Azure Cognitive Search o similar, según la arquitectura OSS). Posteriormente, se ejecutarán consultas de prueba específicas para verificar que el sistema recupere eficazmente el contexto normativo correcto.

Fase 4: Despliegue y Configuración Inicial de la Solución Base.

Establecimiento del Entorno y Despliegue.

Se configurará el entorno de desarrollo necesario y se desplegará la solución open-source de Microsoft. Este proceso se guiará por la documentación oficial del repositorio y se emplearán herramientas como Docker y Azure para establecer la arquitectura RAG fundamental.

Verificación del Funcionamiento Esencial.

Se realizarán pruebas funcionales preliminares utilizando documentos de prueba genéricos. El objetivo será validar la correcta operación del pipeline de recuperación de información y generación de respuestas. Se ajustarán parámetros básicos del modelo de lenguaje (ej. límites de contexto, directrices de estilo) para orientarlo hacia un lenguaje técnico-jurídico.

Fase 5: Evaluación Funcional Inicial del Chatbot.

Pruebas de Funcionalidad con Casos de Uso Reales.

Se diseñará un conjunto de casos de prueba que simulen consultas reales sobre normativas laborales, subsidios y reglamentos internos. Las respuestas generadas por el chatbot serán revisadas manualmente por el equipo del proyecto para evaluar su precisión, coherencia y la correcta fundamentación en el corpus documental.

Documentación de Hallazgos y Definición de Mejoras.

Se registrarán sistemáticamente los resultados de las pruebas, incluyendo

casos de éxito, ejemplos de respuestas satisfactorias, así como cualquier ambigüedad o respuesta incompleta. Esta documentación será la base para identificar y priorizar mejoras futuras (ej. optimización del filtrado, ajuste de la presentación de respuestas).

Presupuesto de Costos – Proyecto RAG para Cajas de Compensación Familiar

Este presupuesto presenta la estimación de costos para desarrollar e implementar un sistema de RAG (Retrieval-Augmented Generation) destinado a facilitar el acceso y comprensión de normativas dentro de las Cajas de Compensación Familiar en Bogotá, Colombia. El proyecto contempla dos fases:

- Fase de Prototipo: Desarrollo inicial con costos centrados en recurso humano por hora.
- Fase de Ambiente Productivo: Despliegue completo con costos mensuales de operación, incluyendo infraestructura, licencias y soporte.

Fase de Prototipo

Duración: 5 meses

Modalidad: Costo por hora

Objetivo: Diseñar y validar un prototipo funcional usando servicios de Azure.

Costos Directos

Recursos Humanos

- Ingeniero de la nube
    - Funciones: Asesoría, implementación y despliegue de un Solution Accelerator
    - Costo: $270.000 COP/hora × 20 horas = $5.400.000 COP
- Total Recursos Humanos: $5.400.000 COP

Servicios Azure (Costo estimado por 5 meses)

- Azure OpenAI Service (GPT-4o): $10.00 USD/mes
- Azure AI Search: $103.85 USD/mes
- Key Vault: $0.03 USD/mes
- Virtual Machines: $4.72 USD/mes

- Virtual Network: $0.40 USD/mes
- Storage Accounts: $5.20 USD/mes
- Otros servicios (Load Balancer, Event Grid, Defender, etc.): $0.00 USD/mes
- Total mensual: $124.20 USD
- Total prototipo: $621 USD
- Total por 5 meses (1 USD → COP 4,322): $2.678.400 COP

Costos Indirectos

- Soporte y mantenimiento básico: Incluido en horas de desarrollo
- Energía e internet desarrolladores: $100.000 COP/mes × 5 meses = $500.000 COP
- Reuniones y comunicación: Sin costo adicional (uso de plataformas institucionales)
- Capacitación básica en Azure: Autogestionada, sin costo adicional

Resumen Fase de Prototipo

- Costo equipo humano: $5.400.000 COP
- Costo servicios Azure: $2.678.400 COP
- Costos indirectos estimados: $500.000 COP

- Costo total prototipo: $8.578.400 COP

Fase de Ambiente Productivo

Duración: Mensual (costos recurrentes)
Objetivo: Operar el sistema RAG en producción con soporte, mantenimiento y escalabilidad asegurada

Costos Directos

Servicios Azure

- Costo mensual estimado: $1.094.731,20 COP Infraestructura Complementaria

(Estos costos difieren del prototipo porque en la fase productivo se necesitan más recursos.)

- Costos relacionados con conectividad, respaldo, y seguridad externa: $800.000 COP

Recursos Humanos

- Operación, monitoreo, soporte, mantenimiento y mejora continua: $38.500.000 COP
- Subtotal Costos Directos: $40.394.731,20 COP

Costos Indirectos

- Servicios públicos (energía, internet): $500.000 COP
- Capacitación continua: $300.000 COP
- Transporte: $200.000 COP
- Varios (papelería, licencias menores, imprevistos): $100.000 COP
- Subtotal Costos Indirectos: $1.100.000 COP

Resumen Fase Productiva (mensual)

- Costo servicios Azure: $1.094.731,20 COP
- Infraestructura complementaria: $800.000 COP
- Costo equipo humano: $38.500.000 COP
- Costos indirectos estimados: $1.100.000 COP
- Costo total mensual estimado: $41.494.731,20 COP

Consideraciones Finales

- Tasa de cambio utilizada: USD 1 = COP $4.322,07.
- Fase prototipo (5 meses): $8.578.400 COP
- Fase producitva costo mensual: $41.494.731,20 COP

Documentación de la ejecución de la metodología

La implementación del prototipo del chatbot RAG se llevó a cabo siguiendo el plan metodológico previamente definido, aprovechando la solución de código abierto de Microsoft como base tecnológica. A continuación, se detalla la ejecución de cada fase:

Fase 1: Establecimiento de Iteraciones Ágiles de Mejora (Ejecución Inicial)

Construcción del Backlog Inicial

Con base en los hallazgos de la Fase 3, se construyó un backlog inicial de mejoras. Este incluyó tareas como la incorporación de un volumen mayor y más diverso de normativa, ajustes en la redacción de los prompts del sistema para refinar el tono de las respuestas, y la exploración de funcionalidades adicionales (ej. opción de descarga de la respuesta en formato PDF.

Definición de Ciclos de Actualización

Se definieron ciclos cortos de actualización, con una previsión mensual, aprovechando la infraestructura cloud y la modularidad del repositorio base para facilitar la implementación ágil de cambios.

Fase 2: Adopción de un Enfoque Adaptable y Centrado en el Usuario (Ejecución Inicial)

Se reconoció desde el inicio que la interacción real con el sistema es fundamental para su perfeccionamiento. Se estableció la premisa de mantener una actitud de mejora continua, priorizando la retroalimentación que se obtendrá progresivamente. Durante las fases iniciales, se buscó prevenir proactivamente errores estructurales comunes, como la confusión de fuentes documentales o la provisión de un contexto insuficiente al modelo, con el objetivo de sentar las bases para una herramienta funcional y confiable.

Se comenzaron a ejecutar los primeros componentes del plan de validación y pruebas:

Ejecución de Pruebas Iniciales (Pre-producción):

Pruebas de Recuperación Normativa.

Se realizaron pruebas unitarias con un conjunto acotado de queries con respuestas conocidas.

Pruebas Funcionales de Generación.

Un subconjunto de respuestas fue evaluado por el equipo según los criterios de exactitud legal, integridad y claridad.

Las pruebas de rendimiento, usabilidad extensiva y casos límite se planificaron para etapas posteriores, una vez el corpus normativo estuviera más consolidado.

Fase 3: Carga y Estructuración del Corpus Normativo Específico

Construcción del Corpus

Se recopiló un conjunto inicial de documentos legales y normativos claves aplicables a las Cajas de Compensación, que incluyó secciones relevantes del Código Sustantivo del Trabajo, la Ley 100, así como ejemplos de reglamentos internos y circulares institucionales. Estos documentos fueron organizados en una estructura de carpetas por temática y se les asignaron metadatos básicos (ej. nombre_documento, fecha_publicacion, entidad_emisora) conforme a los requerimientos del sistema para su posterior indexación.

Carga al Sistema y Validación de Resultados de Indexación

Los documentos seleccionados fueron cargados al sistema utilizando el proceso de ingestión de datos provisto por la solución base, el cual internamente gestiona la indexación (en este caso, apoyado en Azure Cognitive Search). Se realizaron consultas de prueba enfocadas en los documentos recién cargados, validándose que el sistema era capaz de recuperar fragmentos de texto normativo correctos y relevantes para dichas consultas, antes de la generación de la respuesta final.

Fase 4: Despliegue y Configuración Inicial de la Solución

Establecimiento del Entorno y Despliegue

Se configuró el entorno de trabajo y se desplegó la solución open-source proporcionada por Microsoft. Para ello, se siguieron las directrices de la documentación oficial de su repositorio, utilizando Docker para la contenerización de la aplicación y servicios de Azure para su hospedaje y gestión. La arquitectura RAG preconstruida permitió un despliegue ágil sin necesidad de desarrollo de componentes base desde cero.

Verificación del Funcionamiento Esencial

Se realizaron pruebas iniciales con un conjunto de documentos de prueba genéricos. Estas pruebas validaron que el pipeline de recuperación de información y generación de

respuestas operaba correctamente. Se procedió a configurar parámetros básicos del modelo de lenguaje, tales como los límites de tokens para el contexto y prompts iniciales para guiar el estilo de las respuestas hacia un lenguaje técnico-jurídico apropiado.

Fase 5: Evaluación Funcional Inicial del Chatbot

Pruebas Básicas con Usuarios (Equipo de Proyecto)

Se evaluó el funcionamiento del chatbot mediante la formulación de preguntas sobre diversos tipos de normativas laborales, procesos de subsidios y ejemplos de reglamentos internos que habían sido cargados previamente.

Revisión y Documentación de Resultados

Las respuestas generadas fueron revisadas manualmente por el equipo del proyecto. Se documentaron casos de éxito donde la respuesta fue precisa y bien fundamentada, así como ejemplos de respuestas satisfactorias. También se identificaron y registraron algunos casos que presentaron ambigüedad o generaron respuestas incompletas. Estos hallazgos sirvieron para identificar mejoras potenciales a implementar en futuras iteraciones, tales como la necesidad de un mejor filtrado de los segmentos de documentos o una estructura más clara en la presentación de las fuentes.

Link presentación

https://youtu.be/C-ic2i2OYIs?si=tWYtP1Gm_XelH6wm

Conclusiones

El presente trabajo documenta el desarrollo, la implementación y la evaluación funcional inicial de un prototipo de agente virtual (chatbot) basado en la tecnología de Generación Aumentada por Recuperación (RAG). Este prototipo se enfocó en validar la viabilidad técnica de utilizar RAG para mejorar el acceso a la información normativa de las Cajas de Compensación Familiar en Colombia, partiendo de una solución de código abierto (Microsoft RAG OSS).

Aspectos Novedosos:

Se concluye que la tecnología RAG, implementada mediante la adaptación de una solución de código abierto, representa un enfoque técnico viable y prometedor para abordar la complejidad en el acceso a la normativa de las Cajas de Compensación. La principal innovación de este proyecto radicó en la aplicación práctica y la configuración de dicha arquitectura en el contexto específico de la normativa colombiana para estas entidades, demostrando la capacidad de ingestar documentos locales y generar respuestas contextualizadas. Si bien la adaptación cultural exhaustiva de la interfaz de usuario no fue el foco de este prototipo técnico, se reconoce su criticidad, y las pruebas funcionales subrayan la necesidad de considerar la experiencia del usuario final en futuras etapas.

Grado de Cumplimiento de los Objetivos:

El proyecto alcanzó satisfactoriamente los objetivos planteados para la fase de prototipado:

Se implementó un prototipo funcional del chatbot RAG, validando su capacidad técnica para procesar consultas sobre normativa de las Cajas de Compensación Familiar.

- Se desplegó exitosamente la solución de código abierto (Microsoft RAG OSS) como base arquitectónica, utilizando Docker y Azure en un entorno de pruebas.
- Se estructuró e ingestó un corpus documental inicial con normativa específica, verificando la correcta recuperación de información contextual relevante mediante el motor de búsqueda integrado (Azure Cognitive Search).
- Se realizó una evaluación funcional preliminar del prototipo, ejecutando consultas sobre la normativa cargada, analizando la precisión de las respuestas y documentando los hallazgos iniciales, lo que sienta las bases para la mejora continua.

Metodología para la selección y desarrollo de la solución

La metodología adoptada, centrada en la adaptación de una solución de código abierto existente, demostró ser eficiente para la rápida implementación y validación técnica del prototipo. Este enfoque permitió concentrar esfuerzos en la personalización del corpus y en

las pruebas funcionales iniciales, en lugar de en el desarrollo desde cero de la infraestructura RAG. Se concluye que esta estrategia es adecuada para fases tempranas de proyectos de IA donde se busca una validación ágil.

Limitaciones del Proyecto:

Si bien el prototipo fue funcional, esta fase del proyecto presentó las siguientes limitaciones:

- El corpus documental utilizado fue una selección inicial y no abarcó la totalidad de la normativa existente.
- La evaluación funcional se realizó internamente por el equipo del proyecto; no incluyó pruebas con usuarios finales de las Cajas de Compensación, lo que limita las conclusiones sobre usabilidad y experiencia de usuario real.
- No se desarrolló una interfaz de usuario final personalizada; las pruebas se basaron en las capacidades de interacción del sistema OSS.
- Persisten restricciones inherentes a proyectos de IA, como la necesidad de actualización continua de la información, la supervisión de la precisión de las respuestas y la escalabilidad de los costos de procesamiento a medida que aumenta el uso y el volumen de datos.

Proyecciones y Posibilidades Futuras:

La implementación exitosa de este prototipo RAG sienta una base sólida y abre proyecciones significativas. Se concluye que existe un alto potencial para transformar la interacción de los afiliados con las Cajas de Compensación, ofreciendo un acceso más eficiente a la información.

Referencias

Abdellatif, A., et al. (2022). Chatbot implementation challenges. *Journal of Business Tech*, 17(3), 45-67.

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Press.

Alavi, M., & Leidner, D. E. (2001). Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly, 25*(1), 107-136. https://doi.org/10.2307/3250961

Allen, J. F., Byron, D. K., Dzikovska, M., Ferguson, G., Galescu, L., & Stent, A. (2001). Toward conversational human-computer interaction. *AI Magazine, 22*(4), 27-27. https://doi.org/10.1609/aimag.v22i4.1590

Birch, K. (2023, diciembre 19). *How AWS makes data centres more efficient and sustainable*. Bizclik Media Ltd. https://sustainabilitymag.com/net-zero/how-aws-makes-data-centres-more-efficient-and-sustainable

Borgefalk, P., et al. (2023). Legal information retrieval using RAG architecture. *Journal of Artificial Intelligence Research*, 45(2), 123-145.

Brynjolfsson, E., & McAfee, A. (2017). *Machine, platform, crowd: Harnessing our digital future*. W. W. Norton & Company.

Chatbots para el gobierno - IBM watsonx Assistant. (s/f-a). Ibm.com. Recuperado el 27 de marzo de 2025, de https://www.ibm.com/es-es/products/watsonx-assistant/government

Decreto 1377 de 2013. (2013). *Por el cual se reglamenta parcialmente la Ley 1581 de 2012 sobre protección de datos personales*. Presidencia de la República de Colombia. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=53646

De documentos en las entidades públicas, I. de T. de R. S. L. P. la E. y. G. (s/f). *Ingeniería Informática*. Uva.es. Recuperado el 27 de marzo de 2025, de https://uvadoc.uva.es/bitstream/handle/10324/71509/TFM-G1993.pdf;jsessionid=EFC94EDB66554A6732EBE569F4DBC2A2?sequence=1#:~:text=%EF%82%B7%20Fine,completamente%20malinterpretadas%20en%20su%20aplicaci%C3%B3n

Floridi, L., et al. (2021). Ethical guidelines for trustworthy AI. *AI Ethics Journal*, 3(1), 56-78

Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. *Interactions*, *24*(4), 38-42. https://doi.org/10.1145/3085558

HeidiSteen. (s/f). *RAG and generative AI - Azure AI Search*. Microsoft.com. Recuperado el 27 de marzo de 2025, de https://learn.microsoft.com/en-us/azure/search/retrieval-augmented-generation-overview

Hislop, D. (2013). *Knowledge management in organizations*. Oxford University Press.

ISO/IEC 27001. (2022). *Information security, cybersecurity and privacy protection - Information security management systems - Requirements*. International Organization for Standardization.

Lewis, P., Oguz, B., Rinott, R., Riedel, S., & Stoyanov, V. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9450-9462.

Llugsa, D. A. M., & Vaca, J. E. J. (2022). Chatbot una herramienta de atención al cliente en tiempos de COVID-19: Un acercamiento teórico. *Uniandes Episteme, 9*(3), 327-350.

Mercado laboral de Bogotá solo emplea formalmente al 40% de personas en edad de trabajar - Probogotá. (2023, mayo 3). Probogotá. https://www.probogota.org/comunicacion_c/mercado-laboral-de-bogota-solo-emplea-formalmente-al-40-de-personas-en-edad-de-trabajar/

Merritt, R. (2025, enero 31). *What Is Retrieval-Augmented Generation aka RAG*. NVIDIA Blog. https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford University Press.

Piattini, M., & Velthuis, F. (2020). Challenges in managing regulatory compliance in organizations. *Information Systems Journal, 30*(2), 189-210.

Pinzón, M. E., Mojica, L. A., & Jiménez, J. E. (2024). *Seminario de investigación: Chatbot para la automatización de consultas normativas*. Universidad EAN.

(S/f-a). Aoc.cat. Recuperado el 27 de marzo de 2025, de https://www.aoc.cat/es/blog/2024/xatbot-aoc-iagenerativa/#:~:text=La%20AOC%20est%C3%A1%20comprometida%20al,los%20derechos%20de%20los%20usuarios

Servicio al Cliente, IA y Negocios. (s/f). Aivo.co. Recuperado el 27 de marzo de 2025, de https://es.aivo.co/blog

Sharda, R., Delen, D., & Turban, E. (2021). *Analytics, data science, & artificial intelligence: Systems for decision support* (11th ed.). Pearson.

Sirovy, L. (2023, junio 12). *10 métricas clave para evaluar el rendimiento de tu chatbot*. Inbenta. https://www.inbenta.com/es/articles/10-key-metrics-to-evaluate-your-ai-chatbot-performance/

What is retrieval-augmented generation (RAG)? A practical guide. (s/f). K2view.com. Recuperado el 27 de marzo de 2025, de https://www.k2view.com/what-is-retrieval-augmented-generation

Yang, J., Chen, Y.-L., Por, L. Y., & Ku, C. S. (2023). A systematic literature review of information security in chatbots. *Applied Sciences*, *13*(11), 6355. https://doi.org/10.3390/app13116355