**Documento proceso**

**PROYECTO INTEGRADOR**

**Resolución de PQR con IA - EFIGAS**

****

**Profesor:**

**CARLOS ALBERTO RUIZ VILLA**

**Natalia Dané Arboleda Quinchía**

**Nicolás Castro Giraldo**

**Juan Sebastián Giraldo Trujillo**

**Universidad de Caldas**

**Septiembre 2024**

**Propósito del Proyecto**

El propósito del proyecto Resolución de PQR con IA en Efigas es implementar una solución de inteligencia artificial avanzada que permita a los usuarios interactuar de forma intuitiva y fluida con un sistema automatizado para resolver sus dudas, inquietudes y solicitudes relacionadas con los servicios de Efigas. La solución propuesta busca mejorar la experiencia del cliente al reducir el tiempo de respuesta y ofrecer respuestas más precisas, eliminando la necesidad de navegación a través de un menú rígido de opciones. Con este sistema, los usuarios pueden formular preguntas de manera libre y obtener respuestas inmediatas en un frontend conocido para el usuario como lo es whatsApp

Además, el proyecto busca reemplazar el actual chatbot de WhatsApp, basado en un árbol de decisiones, con una plataforma de IA que entiende y procesa las consultas de los usuarios en lenguaje natural. Esto se logra mediante el uso de modelos de lenguaje que analizan documentos de procesos específicos de Efigas, optimizados para proporcionar una interacción más personalizada y eficiente.

**Audiencia**

Este documento de proceso está dirigido a todos los miembros involucrados en el desarrollo e implementación del proyecto de Resolución de PQR con IA para Efigas. La audiencia clave incluye:

* Equipo de Proyecto: Compuesto por los ingenieros de sistemas responsables de la codificación, entrenamiento del modelo y configuración de la base de datos de Chroma. Este grupo utiliza el documento para comprender los objetivos técnicos, los detalles de implementación, las metodologías, y los plazos del proyecto.
* Gerentes de Proyecto: Encargados de supervisar el avance y asegurar que el proyecto cumpla con los objetivos y cronogramas establecidos. Los gerentes de proyecto utilizan este documento para rastrear las fases del proyecto, hitos clave y asignación de recursos.
* Personal de Negocio de Efigas: Incluye personal de soporte y servicio al cliente, quienes necesitan comprender cómo funcionará el sistema de IA y los beneficios específicos para la operación de atención al cliente. Este grupo podrá ver cómo la IA facilitará la interacción con los usuarios finales y contribuirá a la satisfacción del cliente.
* Asesores Académicos: Profesor que proporciona apoyo técnico y metodológico al equipo de proyecto. Se revisará este documento para evaluar la aplicación práctica de las habilidades adquiridas, alineándose con los objetivos de aprendizaje y asegurando que se empleen buenas prácticas.
* Usuarios Finales (Clientes de Efigas): Aunque el documento no está directamente orientado a ellos, el equipo de soporte y desarrollo necesita mantener siempre presente que los usuarios finales son quienes interactuarán con el sistema. Los clientes determinarán el éxito del proyecto mediante su experiencia con el sistema de IA.

Este enfoque colaborativo asegura que todos los interesados comprendan los pasos, objetivos y responsabilidades del proyecto, permitiendo una ejecución clara y alineada con las metas estratégicas y operativas de Efigas

**Fases del Proyecto**

El proyecto de IA de Resolución de PQR con Efigas se divide en fases clave, cada una de las cuales se compone de tareas específicas y objetivos técnicos necesarios para asegurar el éxito del sistema automatizado de atención al cliente. Estas fases son las siguientes:

**1. Recopilación de Datos**: Reunir todos los documentos relevantes de Efigas, incluyendo manuales de procedimientos, preguntas frecuentes, políticas, el protocolo de CHATBOT anterior, documentación de página web útil para las PQR de los usuarios.

* **Identificación de Documentos Relevantes**: Se trabaja en colaboración con el equipo de soporte de Efigas para identificar todos los documentos que el sistema necesita para responder a las preguntas de los usuarios. Estos documentos incluyen:
  + Manuales de procesos internos de Efigas. Se encontraron 11 procesos principales dentro de la atención al cliente: Trámites, Revisión técnica, red nueva, modificación, manual de usuario, GNV, factura, comercio, cliente constructor, brilla, contrato prestación servicios.
  + Documentos de políticas y procedimientos relacionados con los servicios de Efigas.
  + Bases de preguntas frecuentes y respuestas comunes de PQR anteriores.
* **Obtención y Acceso a Documentos**: Una vez identificados, se asegura que todos los documentos estén en formatos accesibles para su procesamiento, en esta caso PDF, que luego pueden transformarse en texto.
* **Actualización y Verificación**: Para evitar desinformación, se verifica que todos los documentos sean actuales y reflejen la información más reciente de Efigas, evitando problemas futuros de precisión en las respuestas del sistema. Comprobando toda esta información con la que se presenta en la página web de la empresa

**2. Limpieza y Preprocesamiento de Datos**: Después de la recopilación de datos, es necesario limpiar y preparar el texto. La limpieza asegura que los documentos se encuentren en un formato adecuado para el procesamiento de lenguaje natural, eliminando cualquier ruido o información irrelevante que pueda afectar el rendimiento del modelo de IA.

* **Eliminación de Ruido y Formato Innecesario**: Se eliminan caracteres especiales, espacios redundantes, encabezados y pie de página que no sean relevantes. También se retiran frases repetitivas que aparecen en múltiples documentos, como “finalizar consulta” o “elige una opción”. Todo esto principalmente sobre el documento de protocolos de chatbot, porque dicho documento estaba lleno de toda la información basada en el árbol de decisiones.
* **Estandarización del Texto**: Para garantizar que todos los documentos tengan una estructura uniforme, se aplica una estandarización de datos. Esto incluye uniformizar formatos de fechas, unidades de medida y términos específicos de Efigas, asegurando que el sistema de IA los reconozca de manera coherente.
* **División en Secciones y Subsecciones**: Los documentos extensos se dividen en secciones y subsecciones lógicas (por ejemplo, Facturación, Revisión Técnica Reglamentaria, etc.) para facilitar el procesamiento y búsqueda. Esto permite al sistema encontrar respuestas más específicas y enfocadas en la sección correcta. A estas subsecciones dentro de la empresa se le conocen como procesos.
* **Revisión de Calidad**: Para confirmar que los datos están listos para el siguiente paso, se revisa la limpieza y estandarización en una muestra de los documentos. Esto garantiza que el contenido se mantenga relevante, claro y esté alineado con las posibles consultas de los usuarios.

**3. Tokenización y Almacenamiento de Embeddings**: Una vez que los datos están limpios, el siguiente paso es tokenizar el texto y convertirlo en **embeddings**, es decir, representaciones vectoriales que capturan el significado de cada fragmento de texto. Este proceso es esencial para habilitar una búsqueda semántica en la base de datos, permitiendo que el sistema comprenda la intención detrás de las preguntas de los usuarios.

* **Tokenización del Texto**: La tokenización divide el texto en fragmentos significativos (palabras, frases o párrafos). Cada fragmento corresponde a una unidad semántica manejable, lo que permite una segmentación precisa del contenido. Esto es especialmente importante en consultas largas o complejas donde cada palabra o conjunto de palabras aporta información importante.
  + **Segmentación en Fragmentos Temáticos**: Se divide el contenido en fragmentos relacionados temáticamente, de modo que cada fragmento represente una parte específica de la información (por ejemplo, opciones de pago, detalles de la revisión técnica, etc.).
* **Creación de Embeddings**: Utilizando un modelo de embeddings cada fragmento de texto se convierte en un vector en un espacio de alta dimensionalidad. Estos embeddings capturan el significado semántico, permitiendo que fragmentos con significados similares estén cerca unos de otros en el espacio vectorial.
  + **Configuración de Modelo de Embeddings**: Se selecciona un modelo de embeddings preentrenado y adecuado para el dominio de preguntas frecuentes y atención al cliente, optimizado para preguntas en lenguaje natural.
  + **Codificación de Fragmentos en Vectores**: Cada fragmento tokenizado se convierte en un vector que representa su contenido semántico. Estos vectores son cruciales para el proceso de búsqueda en la base de datos de Chroma, ya que permiten que el sistema encuentre los fragmentos más relevantes para cada consulta.
* **Almacenamiento en Base de Datos Vectorizada (Chroma)**: Los embeddings generados se almacenan en **Chroma**, la base de datos vectorizada que facilita la búsqueda semántica rápida y eficiente. Cada vector se asocia con metadatos que incluyen el nombre del documento y la página de origen, para que el sistema pueda referenciar la fuente al responder a las preguntas de los usuarios.
  + **Organización de los Embeddings**: Cada fragmento se etiqueta con información adicional (metadatos), como la fuente, la sección y la página, de modo que el sistema pueda ofrecer respuestas contextualizadas y referenciadas.

**4. Entrenamiento del Modelo**: El entrenamiento y ajuste del modelo de IA es esencial para que el sistema de Efigas pueda interpretar correctamente las consultas de los usuarios y proporcionar respuestas pertinentes. En este proyecto, se utiliza un modelo de la familia **GPT**, que ha sido preentrenado para el procesamiento de lenguaje natural, optimizado para comprender consultas complejas y contextuales. Este modelo de lenguaje avanzado no solo reconoce palabras clave, sino que también es capaz de identificar la intención del usuario y asociar su pregunta con los fragmentos de información adecuados en la base de datos de Chroma.

**Selección y Configuración del Modelo de PLN (GPT)**

* **Selección de GPT como Modelo de PLN**:
  + Para el procesamiento de lenguaje natural, se elige un modelo GPT avanzado, capaz de manejar consultas complejas y comprender relaciones semánticas entre palabras en contexto. Esta capacidad es crucial para que el sistema de Efigas pueda responder de manera precisa a preguntas no estructuradas, como las consultas de los usuarios.
  + GPT es compatible con el enfoque de **RAG** (Retrieval-Augmented Generation), permitiendo que el sistema acceda a los fragmentos relevantes en la base de datos de Chroma y los utilice en la generación de respuestas.
* **Integración con Chroma para Búsquedas Semánticas**:
  + La estructura de GPT permite su integración con bases de datos vectorizadas como Chroma, donde se almacenan los embeddings. Esto facilita el proceso de búsqueda semántica, permitiendo que el modelo seleccione rápidamente los fragmentos de texto más relevantes en respuesta a las consultas de los usuarios.

**Preparación de Datos para el Ajuste y Contextualización del Modelo**

**Uso de Ejemplos de Preguntas Reales y Simuladas**:

* + Aunque GPT está preentrenado, se refina aún más para adaptarlo al contexto de Efigas utilizando ejemplos específicos de preguntas reales o simuladas que los clientes podrían hacer. Estos ejemplos abarcan temas como facturación, revisión técnica, políticas de servicio, entre otros.
  + Se asegura que el modelo entienda los temas más consultados de los usuarios de Efigas, proporcionándole acceso a los documentos de soporte, procedimientos y FAQs recopilados en la fase inicial de datos.

**Alineación con la Terminología y Políticas de Efigas**:

* + Para aumentar la relevancia, el modelo se ajusta utilizando un corpus de documentos con terminología y contenido específico de Efigas. Esto permite que el modelo reconozca términos únicos y contextuales de la empresa, mejorando su capacidad de ofrecer respuestas alineadas con los procedimientos y políticas de Efigas.

**5. Implementación de un Sistema de RAG (Retrieval-Augmented Generation)**El sistema RAG es el núcleo que permite al modelo de IA combinar la recuperación de fragmentos relevantes con la generación de respuestas coherentes y específicas. Este enfoque asegura que las respuestas estén basadas en la información real contenida en los documentos, evitando respuestas genéricas y mejorando la relevancia para el usuario.

* **Configuración del Pipeline de RAG**:
  + El sistema RAG se configura para que el proceso de generación de respuestas comience con la recuperación de fragmentos relevantes. Esto significa que, cuando el usuario formula una pregunta, el modelo primero accede a Chroma para encontrar fragmentos de texto que contengan información relacionada con la consulta.
  + En este pipeline, la **búsqueda semántica** permite que la pregunta del usuario se compare con los embeddings almacenados, identificando así los fragmentos más relevantes basados en la similitud semántica.
* **Proceso de Recuperación Inicial**:
  + La recuperación inicial identifica los **5 fragmentos más relevantes** en la base de datos. Estos fragmentos son aquellos cuya similitud con la pregunta del usuario es mayor, calculada mediante un sistema de búsqueda vectorial que mide la proximidad en el espacio de embeddings.
  + Estos fragmentos iniciales sirven como “candidatos” para la generación de respuestas, asegurando que el modelo tenga suficiente contexto para construir una respuesta precisa.
* **Filtrado de Relevancia**:
  + Después de la recuperación inicial, se realiza un filtro adicional para determinar si cada fragmento es verdaderamente relevante para la pregunta. Este filtro revisa la pertinencia de los fragmentos seleccionados en relación con la pregunta específica del usuario.
  + Solo los fragmentos que pasan este filtro de relevancia se utilizan para generar la respuesta. Esto ayuda a evitar respuestas redundantes o basadas en información irrelevante, mejorando la precisión final.
* **Generación de Respuesta**:
  + Con los fragmentos filtrados, el modelo de IA pasa a la fase de generación de respuesta. La generación de la respuesta se realiza con el siguiente código:

response = chain({"input\_documents": docs, "question": question})

* + Este código utiliza los fragmentos filtrados junto con la pregunta del usuario para construir una respuesta natural y coherente, basada en la información relevante.
  + La respuesta generada es una síntesis que incluye la información esencial y relevante que el usuario necesita, manteniendo la claridad y concisión.
* **Manejo de Consultas Sin Fragmentos Relevantes**:
  + En caso de que ningún fragmento pase el filtro de relevancia, el sistema solicita al usuario que reformule su pregunta. Este paso garantiza que el sistema solo ofrezca respuestas cuando tiene información suficiente y relevante.
  + El sistema de IA, al detectar esta situación, puede presentar un mensaje del tipo: “No encontré información suficiente. ¿Podrías reformular tu pregunta?”

**6. Evaluación del Modelo**

En esta fase, el modelo y el sistema RAG se someten a pruebas rigurosas para evaluar su rendimiento y realizar ajustes que garanticen respuestas de alta calidad. La evaluación y optimización continua son esenciales para asegurar que el modelo de IA cumpla con las expectativas de precisión y satisfacción del usuario.

**3.1 Evaluación del Modelo**

* **Pruebas de Precisión y Relevancia**:
  + Se utilizan métricas de precisión y relevancia para evaluar qué tan bien el modelo responde a las consultas de los usuarios. Las métricas incluyen:
    - **Precisión**: Qué tan bien el modelo selecciona fragmentos y genera respuestas relevantes para la consulta.
    - **Recuperación de Información**: Qué tan preciso es el sistema al seleccionar los fragmentos más relevantes en la base de datos Chroma.
    - **Exactitud Semántica**: Qué tan bien la respuesta generada refleja el contenido y la intención del fragmento recuperado.
  + Estas pruebas se ejecutan en una muestra de preguntas de prueba que representan escenarios comunes y excepcionales.
* **Pruebas de Usuario con Escenarios Reales**:
  + Se simulan interacciones con usuarios reales para observar cómo el sistema maneja consultas de distintos niveles de complejidad y diferentes estilos de lenguaje.
  + Los usuarios proporcionan retroalimentación sobre la claridad, relevancia y utilidad de las respuestas generadas.
  + Estas pruebas permiten identificar áreas en las que el modelo puede mejorar su comprensión de la intención del usuario.
* **Métricas de Rendimiento**:
  + **Tiempo de Respuesta**: Se mide cuánto tarda el modelo en responder a una consulta, especialmente importante en entornos de atención al cliente donde los tiempos de espera impactan la satisfacción del usuario.
  + **Tasa de Acertividad**: Porcentaje de respuestas que resuelven correctamente las dudas del usuario sin necesidad de reformular la pregunta.
  + **Tasa de Reformulación**: Indica la frecuencia con la que el sistema solicita al usuario reformular su pregunta debido a falta de fragmentos relevantes. Esto puede indicar áreas donde el sistema necesita mejorar en comprensión o recuperación

**7. Despliegue**: La fase de despliegue es el paso final en el proceso de implementación, donde el sistema de IA pasa de ser un prototipo o modelo en pruebas a un entorno de producción que interactúa directamente con los usuarios de Efigas. En esta fase, se asegura que la infraestructura de soporte sea adecuada y se prepara al equipo de atención al cliente para utilizar y supervisar el sistema.

**4.1 Preparación de la Infraestructura para Despliegue**

* **Configuración de Infraestructura**:
  + Se implementa el sistema en servidores o en una plataforma de nube que soporte la carga de consultas esperada. La infraestructura debe garantizar escalabilidad para manejar el tráfico en horas pico sin demoras.
  + Si se utiliza una solución en la nube, se configuran instancias para mantener la disponibilidad y asegurar la estabilidad del sistema.
* **Seguridad y Privacidad de Datos**:
  + Dado que el sistema manejará consultas de clientes, es fundamental implementar protocolos de seguridad y privacidad de datos. Esto incluye la encriptación de datos y políticas de acceso restringido para proteger la información de los usuarios.
  + Se aseguran medidas de cumplimiento normativo para que el sistema maneje adecuadamente cualquier información sensible de acuerdo con la política de Efigas y las leyes locales de protección de datos.

**4.2 Interacción con Usuarios a través de WhatsApp**

La decisión de utilizar WhatsApp como plataforma principal para la interacción con los usuarios simplifica la implementación del sistema de IA, aprovechando su interfaz preexistente y familiaridad entre los clientes de Efigas.

* **Diseño de la Interacción en WhatsApp:**
  + WhatsApp proporciona una interfaz sencilla e intuitiva que permite a los usuarios interactuar directamente con el sistema de IA sin requerir un aprendizaje adicional.
  + Las consultas se gestionan a través de mensajes de texto, complementados con menús interactivos al final de la consulta como idea para tomar datos de satisfacción del cliente.
* **Integración con la Plataforma de Atención al Cliente de Efigas:**
  + El sistema de IA se conecta con las APIs de WhatsApp Business, lo que permite gestionar consultas en tiempo real, desde preguntas comunes hasta interacciones más complejas.
  + Cuando es necesario, el sistema puede escalar la consulta a un agente humano, si es necesario o si esa es la decisión del cliente.

Este enfoque garantiza una experiencia de usuario fluida y accesible, optimizando la comunicación en una plataforma conocida mientras se mantienen las capacidades avanzadas del sistema de IA para resolver consultas de manera eficiente.