



**Entendimiento del uso la inteligencia artificial generativa (IAG) y su aplicación a un
pequeño proceso empresarial: La Automatización y la Evolución del Trabajo**

Astrid Daniela Giraldo Urrea

Ingeniería Industrial

Modalidad de Práctica

Trabajo de Grado

Orientador

Ph.D. Pablo Andrés Maya Duque

Universidad de Antioquia

Facultad de ingeniería

Pregrado

Medellín

2025

Cita

(Giraldo Urrea, 2025)

Referencia

(Giraldo Urrea, A. 2025). *Entendimiento del uso la inteligencia artificial generativa (IAG) y su aplicación a un pequeño proceso empresarial: La Automatización y la Evolución del Trabajo*. Trabajo de grado. Universidad de Antioquia, Medellín

Estilo APA 7 (2020)



Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: <http://bibliotecadigital.udea.edu.co>

Universidad de Antioquia - www.udea.edu.co

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

Dedicatoria

Dedico este trabajo a mi familia, quien han estado presentes en cada paso que he dado y quienes me han apoyado incondicionalmente

Agradecimientos

Agradezco a los profesores que hicieron parte todos los años de mi proceso en la universidad, gracias por su paciencia, por compartir sus conocimientos, y ser la guía de millones de estudiantes. Además, quisiera agradecer a mis amigos de la universidad por ser cada uno parte de nuestro proceso, tanto profesional como personal.

Tabla de contenido

Resume	8
Abstract	9
Introducción	10
3 objetivos	12
3.1 Objetivo general	12
3.2 Objetivos específicos.....	12
4 Marco teórico	13
6 Metodología	15
7 Resultados	16
7.1 Modelos de Inteligencia Artificial Generativa	16
7.2 Ajuste de modelos de IAG	18
7.2.1 Técnicas de ajuste: Prompt Engineering	19
7.2.2 Técnica de ajuste: Fine-tuning.	20
7.2.3 Técnica de ajuste: Retrieval-Augmented Generation (RAG)	23
7.3 Desarrollo de aplicaciones prácticas	24
7.3.1 Entrenamiento LLM con técnica RAG	25
7.3.2 Limitaciones	27
7.3.3 Desarrollo del código	28
7.4 Pruebas	32
Pregunta Fuera de Contexto	40
Modelo Reentrenado gpt-3.5 turbo	40
Modelo Reentrenado GPT-3.5 turbo	40
8. Conclusiones	41

Lista de tablas

Tabla 1. Modelos de inteligencia artificial generativa, destacando sus principales características, aplicaciones y quien fue su creador. Elaboración propia.	18
---	----

Lista de figuras

Ilustración 1 Recomendaciones de expertos para obtener mejores resultados. Esta figura muestra 6 claves básicas en el momento de usar la técnica Prompt Engineering	19
Ilustración 2 Contraste de prompts menos eficientes y mejorado. Esta figura muestra ejemplos de prompts básico y como podría ser mejorado, para obtener óptimos resultados. Basados en: (Prompt Engineering - OpenAI API, n.d.)	20
Ilustración 3 Flujograma que implica la aplicación de la técnica fine-tuning. Elaboración propia. Basado en: (Fine-Tuning - OpenAI API, n.d.)	22
Ilustración 4 Proceso del funcionamiento de uso de la técnica RAG y un modelo LLM. Elaboración propia. Basado en : (¿Qué Es RAG?: Explicación de La IA de Generación Aumentada Por Recuperación, AWS, n.d.)	24

Siglas, acrónimos y abreviaturas

IAG	Inteligencia artificial generativa
PYMES	Microempresas, pequeñas y medianas empresas
GAN	Redes Generativas Adversariales
VAE	Decodificadores variacionales
LLM	Modelos de Lenguaje Grande
MMLU	Compresión masiva del lenguaje multitarea
RAG	Retrieval-Augmented Generation

Resume

Este proyecto tiene como objetivo explorar como la IA generativa puede optimizar pequeños procesos empresariales y analizar cuáles son los entornos en que mejor se desempeña. Además, se estudia las diversas técnicas de entrenamiento para estos modelos generativos. La metodología incluyó la revisión de la literatura, la exploración de la forma como personalizar un modelo IAG y una prueba experimental enfocada en abordar necesidades específicas de la población cafetera en Colombia.

La fase experimental consistió en desarrollar y entrenar un modelo basado en investigaciones realizadas por la Federación Nacional de Cafeteros. Este modelo procesa información relacionada con la evolución y manejo de cultivos, así como la implementación de tecnologías emergentes para optimizar la producción y el desarrollo del café. Al ser entrenado con datos específicos, el modelo demostró su capacidad para responder a preguntas de los usuarios basándose en las investigaciones analizadas, ofreciendo soluciones prácticas y contextualmente relevantes para el sector cafetero. Esta capacidad permite ofrecer soluciones prácticas al sector, apoyando decisiones basadas en datos e investigaciones.

Finalmente, se confirmó la utilidad de personalizar modelos de IA generativa para aplicaciones específicas.

Palabras clave: proyecto de grados, inteligencia artificial generativa, rag

Abstract

This project aims to explore how generative AI can optimize small business processes and analyze the environments where it performs best. Additionally, various training techniques for these generative models are studied. The methodology included a literature review, the exploration of how to customize a generative AI model, and an experimental test focused on addressing the specific needs of Colombia's coffee-growing population.

The experimental phase involved developing and training a model based on research conducted by the National Federation of Coffee Growers. This model processes information related to the evolution and management of crops, as well as the implementation of emerging technologies to optimize coffee production and development. After being trained with specific data, the model demonstrated its ability to answer user questions based on the analyzed research, offering practical and contextually relevant solutions for the coffee sector.

This capability enables the model to provide practical solutions to the industry, supporting decisions based on data and research. Finally, the usefulness of customizing generative AI models for specific applications was confirmed.

Keywords: degree project, generative artificial intelligence, retrieval-augmented generation

Introducción

En la actualidad, la revolución digital ha marcado un antes y un después en la manera en que las empresas operan, incitando la adopción de tecnologías emergentes para seguir siendo competitivas en un entorno de constante evolución. Entre estas tecnologías la inteligencia artificial se está fortaleciendo como una herramienta importante, no solo por su capacidad para automatizar tareas y procesar grandes volúmenes de información, sino también por su potencial para transformar los procesos empresariales. La inteligencia artificial se está convirtiendo en una de las tecnologías más polémicas de nuestro tiempo, debido al potencial para cambiar el modo de operar los procesos y al mismo tiempo nuestro entorno social(*El Estado de La IA En 2022 y El Balance de Media Década / McKinsey*, n.d.). En el amplio campo de la inteligencia artificial, La IAG ha emergido como una rama innovadora, debido a que es capaz de generar contenido original de diversas formas, como textos, imágenes, música y datos sintéticos, lo que es clave para la innovación de las organizaciones.

La IAG abre una variedad de oportunidades para las empresas que buscan no solo optimizar sus procesos, sino también explorar nuevos modelos de negocios y mejorar las experiencias con los clientes. Sin embargo, a pesar de su potencial, la adopción de la IAG en empresas de menor escala sigue siendo limitada en algunos países. En Colombia, la adopción de la IAG en pequeñas y medianas empresas (PYMES) aún se encuentra en una fase débil, aunque el interés y la inversión en tecnologías está creciendo rápidamente debido a la necesidad de innovación.(*Colombia Avanza En La Adopción de La Inteligencia Artificial Generativa - Fedesoft*, n.d.).Según Daniel Verswyvel, gerente general de Microsoft Colombia, un estudio de la compañía reveló que, de las 45 regiones de Colombia el 77% de las pymes quieren empezar a usar la inteligencia artificial en sus negocios, pero solo el 46% de ellas se sienten capacitadas para hacerlo(“*Más de 77% de Las Pyme En Colombia Quieren Empezar a Usar Inteligencia Artificial,*” n.d.) Esto se debe en gran parte, a las restricciones en recursos técnicos, así como la comprensión de integrar estas tecnologías de manera efectiva en sus procesos.

El presente proyecto tiene como objetivo principal explorar y comprender cómo la inteligencia artificial generativa (IAG) puede aplicarse de manera efectiva en procesos específicos dentro de las organizaciones, con un enfoque particular en las pequeñas y medianas empresas

(PYMES). A través de un análisis del estado actual de la IAG y su potencial en el ámbito empresarial, se buscó identificar no solo las aplicaciones más adecuadas, sino también las limitaciones y explorar formas de personalizar un modelo de inteligencia artificial generativa.

Además, el proyecto incluye un componente práctico que consistió en la implementación y prueba de un modelo IAG, usando la técnica RAG, técnica que permite la recuperación de información proporcionada al modelo y esta información fue sobre las investigaciones realizadas por la Federación Nacional de Cafeteros.

Este proyecto busca proporcionar un marco práctico y accesible para la implementación de la IAG en procesos empresariales. La aplicación de la IAG en pequeñas empresas no solo tiene la oportunidad de mejorar la eficiencia operativa y reducir costos, sino también nuevas oportunidades para la innovación en productos y servicios.

Además, la justificación de este proyecto se sustenta en la necesidad de ampliar el conocimiento sobre la personalización y adaptar los modelos de IAG para resolver tareas o procesos concretos.

3 objetivos

3.1 Objetivo general

Entender y demostrar cómo la inteligencia artificial generativa (IAG) puede ser aplicada en procesos dentro de las organizaciones.

3.2 Objetivos específicos

- Analizar el estado actual de la IA generativa y su potencial en procesos empresariales.
- Identificar las aplicaciones adecuadas de la inteligencia generativa, las limitaciones y casos no adecuados.
- Identificar y seleccionar pequeño proceso empresarial adecuados para la aplicación de IA generativa.
- Implementar y probar la IA generativa en un proceso seleccionado.
- Medir y analizar el impacto de la implementación de la IA.

4 Marco teórico

La inteligencia artificial generativa se refiere a la capacidad de los algoritmos de aprendizaje automático crear contenido nuevo a partir de datos existentes como texto, imágenes, música y datos sintéticos, todo ello es posible mediante el uso de modelos como; las Redes Generativas Adversariales (GANs), los Modelos de Lenguaje Grande (LLMs) y los Decodificadores variacionales (VAEs). Las Redes Generativas Adversariales (GANs) introducidas por Ian Goodfellow, consisten en dos redes neuronales que compiten entre sí para generar datos realistas, permitiendo generar imágenes y otro tipo de contenido(*de la Torre, 2023*). Mientras que los modelos de lenguaje de gran tamaño (LLM), consiste en un codificador y un decodificador con capacidad de autoatención para procesar grandes cantidades de textos, aprendiendo patrones y relaciones entre palabras en diferentes contextos; GPT-4 (que sustenta ChatGPT) y LaMDA (el modelo detrás de Gemini son ejemplos de LLM (*¿What is LLM? - Large Language Models Explained - AWS, s. f.*) y los modelos generativos VAE se utilizan para generar nuevos datos en forma de variaciones de los datos de entrada con los que se entrenan, es decir se usa para la generación de datos sintéticos(*What Is a Variational Autoencoder? | IBM, n.d.*).

La optimización de procesos empresariales es una estrategia clave para mejorar la eficiencia y reducir costos, especialmente en las pequeñas y medianas empresas (PYMES). Técnicas como la automatización, la mejora continua y el lean management son fundamentales para este propósito. La IA generativa puede desempeñar un papel crucial en este ámbito al automatizar tareas repetitivas, personalizar servicios, generar código, analizar errores, y evaluar el rendimiento. Para aplicar efectivamente la IA generativa en pequeños procesos empresariales, es necesario identificar y seleccionar los procesos adecuados, desarrollar y entrenar modelos específicos, e implementar pruebas piloto para evaluar su rendimiento. Además, es crucial medir y analizar el impacto de la implementación mediante la recopilación de datos antes y después, utilizando métricas de rendimiento claras (*Brynjolfsson & McAfee, 2014*)

Para seleccionar el correcto proceso para automatización on inteligencia artificial generativa, los expertos recomiendan múltiples estrategias. Un importante enfoque que se destaca es evaluar la

relación entre las capacidades, características del modelo de IAG y las necesidades del negocio. Por ejemplo, según la consultora estratégica global McKinsey and Company enfatiza que los diferentes tipos de modelo IAG (como la generación de texto o resúmenes de datos) se deben ajustar a casos de uso específico dentro de la organización como es la atención al cliente (*A CIO and CTO Technology Guide to Generative AI / McKinsey*, n.d.).

En este sentido, algunos expertos han señalado que la inteligencia artificial generativa (IAG) tiene gran potencial para ser aplicada en algunos procesos, en particular se destaca la generación de contenido basados en texto siendo útil en áreas como; la descripción de productos, la generación de informes, además de la implementación de asistentes virtuales, que pueden generar respuestas personalizadas en tiempo real, por otro lado, también se encuentra los procesos relacionadas con el diseño creativo el desarrollo de software, la simulación y modelado. Sin embargo, La Inteligencia Artificial Generativa (IAG) cuenta con limitación para ser aplicada en algunos procesos, especialmente en procesos que demandan precisión extrema como la toma de decisiones en tiempo real o una toma de decisiones crítica. Además, la interpretación de datos complejos particularmente en tareas que requieran un alto nivel razonamiento, como la resolución de problemas matemáticos avanzado. Así mismo en áreas donde se necesite la interacción altamente emocional o juicio humano.

Uno de los aspectos más relevantes en el contexto de la inteligencia artificial generativa en las organizaciones es el uso de conjuntos de datos de entrenamiento. Los datos de entrenamiento para la inteligencia artificial generativa se refieren a la colección de ejemplos o conjuntos de datos que se utilizan para "enseñar" al modelo de IA cómo generar salidas, como texto, imágenes, audio o video (*What Is AI Training Data? Explore AI Training Datasets & Providers / Datarade*, n.d.). Es fundamental seleccionar cuidadosamente los datos con los cuales se quiere entrenar el modelo con el fin de ajustar el modelo a las necesidades esto permite que el modelo genere contenido nuevo que se adapte a la complejidad, el estilo y la estructura de los datos de entrada, además de evitar la generación de contenido sesgado (*Generative AI Data in '25: Importance & 7 Methods*, n.d.).

6 Metodología

La metodología adoptada para el siguiente proyecto es un enfoque mixto, debido a que combina un análisis cualitativo por medio de la revisión de la literatura y un enfoque cuantitativo por medio de la evaluación del modelo de IAG. Para el desarrollo del proyecto se presentaron las siguientes etapas:

La revisión de la literatura se realizó usando diferentes fuentes en la web, como también base de datos académico entre ellas Google scholar y Science Research, donde se exploró múltiples artículos sobre la inteligencia artificial generativa.

La etapa de exploración consistió en la personalización de modelos IAG, para ello se exploraron metodologías de entrenamiento de modelos de inteligencia artificial generativa como el ajuste fino (fine-tuning), la ingeniería de prompt (prompt Engineering) y Retrieval-Augmented Generation (RAG) por medio de estudios prácticos e información especializada en estos temas, permitió identificar las mejores prácticas al momento de personalizar un modelo IAG según las necesidades específicas.

Para finalizar se encuentra la fase experimental, la cual consistió en desarrollar y entrenar un modelo basado en investigaciones realizadas por la Federación Nacional de Cafeteros. Este modelo usando la técnica RAG, procesa información relacionada con la evolución y manejo de cultivos, así como la implementación de tecnologías emergentes para optimizar la producción y el desarrollo del café. Además, el modelo creado sirve como un asistente para las personas interesadas en obtener conocimiento sobre estas nuevas investigaciones, permitiéndoles realizar preguntas específicas sin necesidad de leer cada documento en su totalidad, ya que el modelo proporciona respuestas precisas basadas en la información contenida en las investigaciones

7 Resultados

7.1 Modelos de Inteligencia Artificial Generativa

Para comenzar, es importante destacar algunos de los distintos modelos existentes y conocer sus características distintivas, lo cual es clave para entender cómo se pueden personalizar los modelos de inteligencia artificial generativa. En este sentido, la Tabla 1 expone una comparación de diferentes modelos de inteligencia artificial generativa, proporcionando una visión clara de sus particularidades y posibles aplicaciones

Modelo	Características	Principales aplicaciones	Creadores
Gpt-4	<ul style="list-style-type: none"> -Modelos lingüístico extenso (LLM) -Principalmente construye texto a partir de datos de entrenamiento. -Reconocido por la capacidad de comprender y generar texto coherente. -GPT-4 no tiene por sí mismo la capacidad de generar imágenes. -No tiene la capacidad de identificar qué conceptos son correctos o erróneos. -Uno de los problemas que aún continúa es que inventa datos. -No se recomienda usar para tareas que se requiere precisión. 	<ul style="list-style-type: none"> - Redacción de correos electrónicos - Descripción de productos - Chatbots - Asistente de redacción. - Aprender idiomas - escribir subtítulos - generación de código 	Open AI

<p>Llama 2</p>	<ul style="list-style-type: none"> -El modelo lingüístico -Modelo superior a llama 1 -Llama 2 se entrenó con fuentes de datos en línea disponibles públicamente. -Los modelos de Llama 2 se entrenan con 2 billones de tokens y tienen el doble de longitud de contexto que Llama 1. Mejor rendimiento en lenguaje inglés. 	<ul style="list-style-type: none"> - Creación de aplicaciones IA - Generar contenido escrito creativo - Resumen automático de texto - Asistente de investigación (analizar y procesar grandes conjuntos de datos científicos) - Atención al cliente y chatbots - Traducción de idiomas 	<p>Meta AI</p>
<p>Gpt-3.5-turbo</p>	<ul style="list-style-type: none"> - Modelos lingüístico extenso (LLM) - Es más rápido y económico que GPT-4. - Puede mantener conversaciones largas mientras recuerda el contexto previo - Admite múltiples idiomas - entrenado con un conjunto de datos que incluye libros, artículos, sitios web y otras fuentes hasta septiembre de 2021. - generar información incorrecta o inventada si no tiene datos suficientes para responder - puede olvidar información si se supera el límite de tokens. 	<ul style="list-style-type: none"> - Asistentes virtuales - Generación de contenido - Análisis de datos de texto - Genera ideas, historias o contenido creativo 	<p>Open AI</p>

Gemini ultra	- Modelos IA multimodal	- Ayuda a optimizar ventas, en generación de propuestas y campañas de marketing	Google
	- Con una calificación del 90%, experto en la MMLU.	- Adapta los mensajes según el cliente.	
	- Combinaciones de 57 materias.	- Crea un resumen a partir de la extracción de información útil de documentos.	
	- Puede generar texto, código, imágenes, audio y video.	- Personaliza imágenes - Toma decisiones basados en datos. área de recursos humanos (descripción de trabajos y capacitaciones)	

Tabla 1. Modelos de inteligencia artificial generativa, destacando sus principales características, aplicaciones y quien fue su creador. Elaboración propia.

7.2 Ajuste de modelos de IAG

La personalización de modelos de inteligencia artificial generativa implica adaptar estos modelos pre entrenados a usos específicos o tareas. Con el objetivo de lograr obtener el potencial que estas tecnologías pueden brindar existen diferentes técnicas de ajustes que ayudan a que el modelo pueda generar respuestas más precisas y relevantes, además minimizar errores o problemas de alucinación. A continuación, algunos ejemplos de las técnicas más usadas.

7.2.1 Técnicas de ajuste: Prompt Engineering

Esta técnica consiste en dar instrucciones con la finalidad de crear datos relevantes y de alta calidad. ¿Cómo se realiza? En la ingeniería de prompt se eligen formatos, frases, palabras y símbolos que ayudan a crear múltiples textos de entrada en lenguaje natural donde se solicita a la IAG que realice una tarea específica. Es importante resaltar que, aunque los modelos de IAG son muy potentes no todos los tipos de prompt pueden generar un resultado útil.

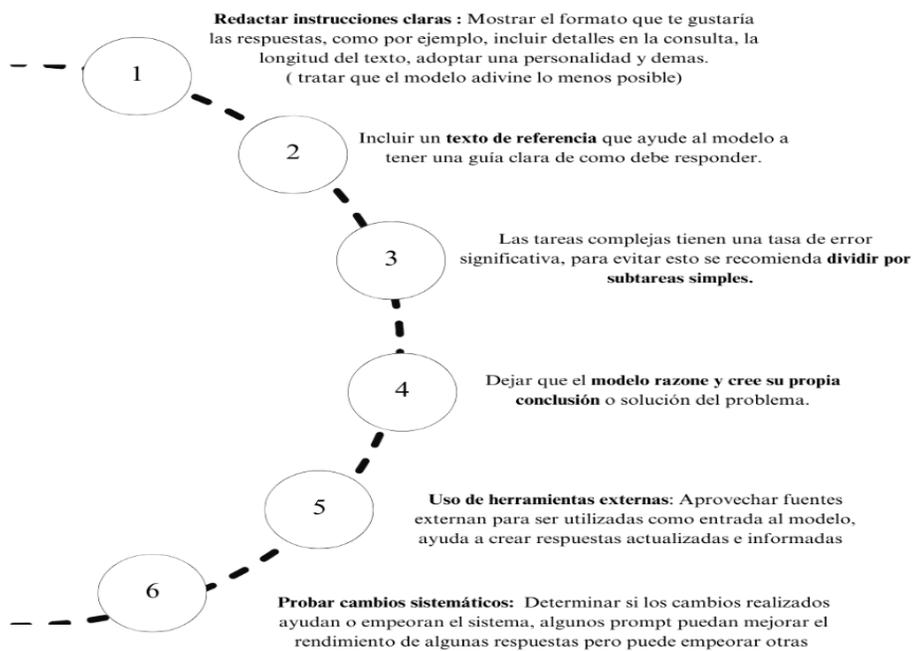


Ilustración 1 Recomendaciones de expertos para obtener mejores resultados. Esta figura muestra 6 claves básicas en el momento de usar la técnica Prompt Engineering

Para crear prompt óptimos, se requiere comunicar instrucciones claras que incluya tanto el contexto como el alcance. Además, es importante emplear expresión creativa para perfeccionar las indicaciones. Por este motivo muchos expertos dan algunas recomendaciones para ser aplicada al momento de construir texto de entrada que ayude a obtener mejores resultados en la interacción de modelos. Algunos de estas recomendaciones se pueden evidenciar en ilustración 1

A continuación, La ilustración 2, logra mostrar un contraste entre un prompt menos eficiente a un prompt mejorado. Cabe resaltar que este es un ejemplo de los más simple en términos de prompts, por otro lado, si se desea ampliar la información acerca de las tácticas y ejemplo más avanzados estos se pueden obtener en la plataforma de open IA.(*Prompt Engineering - OpenAI API*, n.d.)

Worse	Better
How do I add numbers in Excel?	How do I add up a row of dollar amounts in Excel? I want to do this automatically for a whole sheet of rows with all the totals ending up on the right in a column called "Total".
Who's president?	Who was the president of Mexico in 2021, and how frequently are elections held?
Write code to calculate the Fibonacci sequence.	Write a TypeScript function to efficiently calculate the Fibonacci sequence. Comment the code liberally to explain what each piece does and why it's written that way.
Summarize the meeting notes.	Summarize the meeting notes in a single paragraph. Then write a markdown list of the speakers and each of their key points. Finally, list the next steps or action items suggested by the speakers, if any.

Activar Windows
Ve a Configuración para act

Ilustración 2 Contraste de prompts menos eficientes y mejorado. Esta figura muestra ejemplos de prompts básico y como podría ser mejorado, para obtener óptimos resultados. Basados en: (Prompt Engineering - OpenAI API, n.d.)

7.2.2 Técnica de ajuste: Fine-tuning.

El fine-tuning es el proceso de adaptar un modelo de inteligencia artificial generativa (IAG) previamente entrenado para realizar tareas específicas, además es una de las técnicas fundamentales del Deep Learning. Este enfoque se ha convertido esencial para perfeccionar las capacidades de los modelos previamente entrenados a través de la API, logrando un ajuste de manera óptima a nuevas tareas.

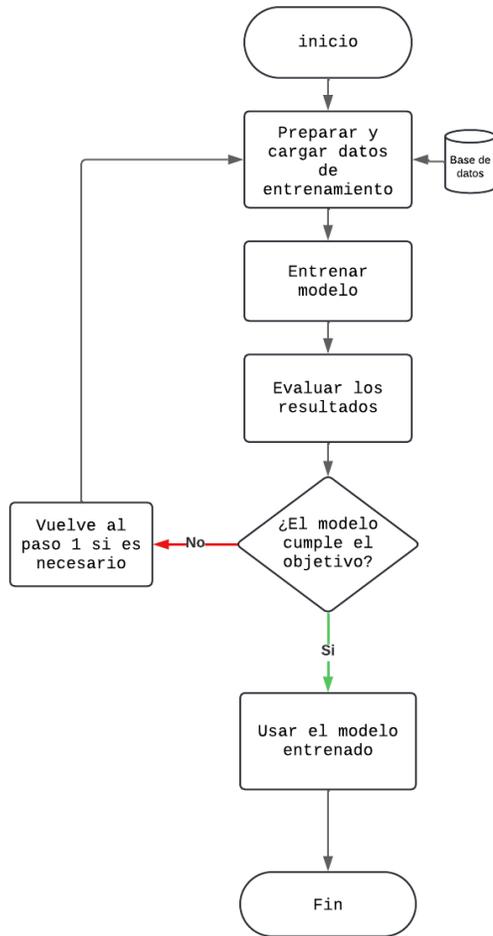
En este contexto, los modelos de lenguaje de gran escala (LLM) son entrenados inicialmente utilizando grandes corpus de datos mediante aprendizaje autosupervisado. Ejemplos destacados de modelos autorregresivos como GPT de OpenAI, Gemini y Llama de Meta están

diseñados para predecir la palabra siguiente en una secuencia, iterando este proceso hasta completar el texto. (*¿Qué Es El Fine-Tuning? | IBM, n.d.*)

El proceso comienza con la selección de un modelo IAG previamente entrenado. A partir de ahí, se realiza un reentrenamiento con un conjunto de datos específicos de la tarea a desempeñar. Estos datos deben incluir preguntas, respuestas relevantes y debe estar directamente relacionados con la tarea que el modelo va a desempeñar, además, los datos deben contener etiquetas de manera clara, precisa y deben estar en formato adecuado para el modelo, como texto tokenizado. Es importante resaltar que la cantidad de datos con las cuales se reentrena el modelo deben ser suficientes y diversos para evitar sesgos al momento que el modelo responda las preguntas solicitadas. (Szép et al., 2024)

Según Open IA para el ajuste de un modelo se requiere que se proporcionen al menos entre 50 y 100 ejemplos de entrenamiento gpt-4o-mini y gpt-3.5-turbo, pero la cantidad correcta varía en gran medida según el caso de uso exacto. (*Fine-Tuning - OpenAI API, n.d.*)

El proceso de implementación del fine-tuning, como se ilustra en la *Figura 3*, que presenta el flujograma que implica la aplicación de la técnica fine-tuning, incluye la evaluación de los resultados obtenidos tras aplicar la técnica. Esta evaluación se realiza comparando el rendimiento del modelo ajustado con el de un modelo base o sin ajustar, utilizando métricas específicas para la tarea, como la exactitud, la precisión o la puntuación. (Howard & Ruder, n.d.)



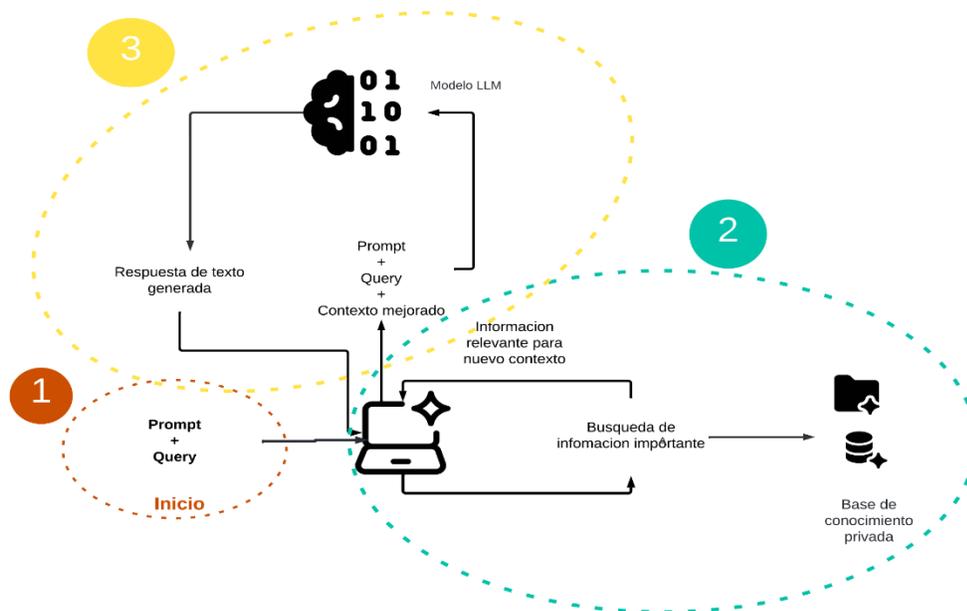
*Ilustración 3*Flujograma que implica la aplicación de la técnica fine-tuning. Elaboración propia. Basado en:(Fine-Tuning - OpenAI API, n.d.)

Para saber cuándo es apropiado usar la técnica de fine-tuning en un modelo IAG, Los expertos recomiendan iniciar usando la técnica de prompt debido a múltiples factores. En primer lugar, en algunas tareas con las indicaciones correctas los modelos pueden mejorar su rendimiento significativamente, evitando los ajustes adicionales. Seguidamente, las iteraciones sobre indicaciones tienen un ciclo de retroalimentación más rápido que usando fine-tuning puesto que requiere crear bases de datos y ejecutar entrenamientos. Por último, se ha demostrado que combinando prompt con la técnica de fine-tuning se han obtenido mejores resultados(*Fine-Tuning - OpenAI API, n.d.*).

Cuando se determina que el fine-tuning es el recurso adecuado a utilizar para entrenar los modelos de IAG, el siguiente paso es preparar los datos para ser usados. Por lo tanto, se debe crear un conjunto de conversaciones o datos similares a los que posteriormente se van a solicitar al modelo, así mismo el conjunto de datos debe contener la misma estructura de la API de finalización de chat, es decir donde cada mensaje tiene un rol y contenido.

7.2.3 Técnica de ajuste: Retrieval-Augmented Generation (RAG)

Es una técnica creada para aumentar la precisión y confiabilidad de los modelos de inteligencia artificial con datos obtenidos de múltiples fuentes externas, es decir los modelos de lenguaje de gran tamaño LLM son entrenados con enormes volúmenes de datos y usan millones de parámetros para responder preguntas por un usuario, usando RAG estas capacidades se pueden extender a información específica como las bases de datos y conocimientos internos de una organización. ¿Cómo funciona? se ingresa un componente de recuperación de información que se utiliza en el texto de entrada del usuario para extraer en primer lugar la información del nuevo origen de datos, seguidamente la consulta del usuario y la información relevante se entrega al modelo LLM y finalmente el modelo LLM utiliza el nuevo conocimiento adquirido y los datos de entrenamiento para crear y mejorar las respuestas (*¿Qué Es RAG?: Explicación de La IA de Generación Aumentada Por Recuperación, AWS, n.d.*). Este proceso se ilustra de manera más clara en la **Figura 4**, que muestra el funcionamiento y uso de la técnica RAG en combinación con un modelo LLM



*Ilustración 4*Proceso del funcionamiento de uso de la técnica RAG y un modelo LLM.Elaboración propia. Basado en : (¿Qué Es RAG?: Explicación de La IA de Generación Aumentada Por Recuperación, AWS, n.d.)

La implementación de este tipo de técnicas (RAG) para el ajuste de modelos de IAG se han considerado beneficiosos para una organización debido a múltiples características. Ante todo, la implementación de esta técnica de ajuste es rentable para ingresar nuevos datos en el LLM. Por otro lado, el uso de RAG permite a los usuarios proporcionar últimas investigaciones e información relevante de las, asimismo, permite conectar el LLM directamente a sitios web, redes sociales, o fuentes de información que se actualizan constantemente, logrando así obtener información más reciente a los usuarios.

7.3 Desarrollo de aplicaciones prácticas

Para este proyecto se decidió abordar un tema importante para la población cafetera de Colombia, tomando como base las investigaciones realizadas por la Federación Nacional de Cafeteros, la cual se creó para mejorarla producción y desarrollo sostenible del café en el país.

Se inicia por revisar los informes y documentos técnicos más recientes que esta organización emite, los cuales se centran en la evolución de prácticas de cultivo, manejo de plagas y usos de tecnologías emergentes para mejorar la calidad del café colombiano.

Luego de la selección de la documentación actual se inicia con la creación de modelo y la implementación de la técnica RAG.

7.3.1 Entrenamiento LLM con técnica RAG

Se seleccionó la técnica RAG porque puede producir texto fluido basado en los documentos proporcionados por la compañía o persona. Para la precisión y contextualización de las respuestas no es necesario reentrenar el modelo con grandes volúmenes de datos. Sin embargo, esto no quiere decir que no pueda recibir grandes cantidades de información, además este modelo también permite acceder a fuentes externas en tiempo real.

Tecnología empleada. Se detalla las tecnologías que fueron utilizadas al momento de desarrollar el modelo y las que fueron claves para llevar a cabo dicho proyecto.

- **Google Colab:** es una plataforma gratuita que permite programar y ejecutar código de Python, además se integra con servicios de Google drive, y su entorno es amigable para llevar una descripción y creación del proyecto.



- **Google drive** es una herramienta que permite almacenar, acceder, crear y compartir archivos. Esta herramienta se usó para almacenar el block de Google Colab y los documentos en pdf que fueron creados para entrenar el modelo.



- **Python:** Es un lenguaje de programación utilizado para diferentes tipos de aplicaciones como desarrollo web, software, ciencia de datos y machine learning. Es un lenguajes eficiente y fácil de aprender, además cuenta con numerosas bibliotecas diseñadas para facilitar muchos tipos de implementaciones, especialmente la técnica RAG y es compatibilidad con modelos pre entrenados.



- **LangChain** es un marco de código abierto diseñado para crear aplicaciones basadas en modelos de lenguaje de gran tamaño (LLM). Ofrece herramientas y abstracciones que facilitan la personalización, mejoran la precisión y aumentan la relevancia de la información generada por estos modelos. Durante el desarrollo y aplicación de la técnica RAG, LangChain desempeñó un papel clave al proporcionar diversos componentes fundamentales y útiles para su implementación (*What Is LangChain? - LangChain Explained - AWS, n.d.*).



- **Api OpenAi** (interfaz de programación de aplicaciones) esta herramienta permite acceder a los diferentes modelos de aprendizaje automáticos pre entrenados, permitiendo que estos puedan ser personalizados para usos específicos. Para el desarrollo del proyecto se debió adquirir el api para poder acceder el modelo **GPT-3-turbo**. Se elige este modelo porque es significativamente más económico y ofrece las mismas funcionalidades (como la velocidad de respuesta y la capacidad para recuperación de datos) en comparación a un modelo más actualizado.



7.3.2 Limitaciones

Durante el desarrollo del proyecto, se presentaron una serie de limitaciones, lo que ha afectado los tiempos y la implementación del modelo usando la técnica RAG.

- **Procesamiento de datos:** Google Colab cuenta con espacio limitado. Esto ocasiona que, al trabajar con documentos extensos, las ejecuciones se detengan.
- Los documentos utilizados para entrenar el modelo deben estar lo más limpios posible. Esto significa que, si los documentos contienen imágenes grandes o elementos visuales que generan ruido, estos deben ser eliminados. Para lograrlo, se recomienda usar **herramientas especializadas en extracción de texto**. Durante la investigación, se identificaron varias opciones como **Docugame** y **Docparser**, los cuales tienen la capacidad de extraer texto con alta precisión, pero presentan un costo elevado. Debido a restricciones presupuestarias, se optó por usar **Adobe Acrobat** como herramienta para limpiar los documentos. Además, se recomienda que los documentos estén formateados con una sola columna. Si los documentos contienen dos columnas, esto puede generar problemas en la extracción de texto y afectar la calidad del preprocesamiento, ya que las herramientas de extracción

pueden interpretar incorrectamente el flujo del texto. Por ello, es importante ajustar el formato de los documentos a una sola columna antes de utilizarlos para entrenar el modelo.

- Librerías de gran tamaño que Google Colab tenía dificultades para procesar, junto con las actualizaciones de LangChain diseñadas para distintos tipos de modelos de IA generativa.
- Las versiones de las librerías y herramientas deben ser compatibles para que no existan problemas entre las diferentes versiones, debido a ello esto creó un pequeño entorpecimiento de los tiempos propuestos.
- Los créditos del api de OpenIA se agotaban.

7.3.3 Desarrollo del código

El modelo desarrollado se basa en el uso de la técnica RAG para crear un modelo que sirve como asistente virtual, el cual consiste en la recuperación de información, es decir este asistente virtual genera respuestas precisas y contextualizadas sobre los documentos entregados al modelo. para la ampliación de este modelo puedes consultar la siguiente referencia (*Astridgiraldo/Proyecto_Rag: Modelo RAG*, n.d.)El modelo se encuentra en un repositorio de la plataforma GitHub

para este proyecto se usaron las investigación realizadas por la federación nacional de cafeteros, las investigaciones son; Vista de Maíz y frijol biofortificados intercalados con café(*Vista de Maíz y Frijol Biofortificados Intercalados Con Café | Avances Técnicos Cenicafé*, n.d.), Software para la trazabilidad de los procesos de producción del café a través de tecnologías blockchain y Análisis de las estrategias de producción de café sostenible para el fortalecimiento del empalme generacional en el municipio de Aguadas (Caldas)(*Vista de Análisis de Las Estrategias de Producción de Café Sostenible Para El Fortalecimiento Del Empalme Generacional En El Municipio de Aguadas (Caldas)*, n.d.)

Estas investigaciones al entregarlas al modelo y ser reentrenado el usuario puede realizar preguntas sobre estos documentos y este modelo genera respuestas precisas basadas en la información proporcionada.

- **Preparación (Descargas)** Se llevó a cabo la descarga de las librerías necesarias para el desarrollo del modelo y la implementación de la técnica RAG. Para ello, se requirieron las siguientes librerías:

```

v Descargas

[ ] %%capture
!pip install langchain pypdf openai chromadb tiktoken

[ ] %%capture
!pip install -U langchain-community

[ ] %%capture
!pip install PyPDF2

[ ] %%capture
!pip install openai

[ ] %%capture
!pip install --upgrade openai
    
```

- **Documentos**

En este apartado se llevó a cabo la carga de la documentación que se utilizaría para reentrenar el modelo. Como se mencionó anteriormente, los documentos deben estar lo más limpios

```

import os
from langchain.document_loaders import PyPDFLoader
from google.colab import drive

# Directorio donde están almacenados tus PDFs en Google Drive
pdf_directory = '/content/drive/MyDrive/prueba y documentos ' # Cambia esto a la ruta correcta

# Obtener la lista de archivos PDF en el directorio
pdf_files = [f for f in os.listdir(pdf_directory) if f.endswith('.pdf')]

ml_papers = []

# Cargar cada archivo PDF
for i, pdf_file in enumerate(pdf_files):
    file_path = os.path.join(pdf_directory, pdf_file)

    # Cargar el PDF con PyPDFLoader
    loader = PyPDFLoader(file_path)
    documents = loader.load()

    ml_papers.extend(documents)
    print(f'Cargado {pdf_file}')

# Ver el número de documentos cargados
print(f'Se han cargado {len(ml_papers)} paginas.')
    
```

posible para garantizar un procesamiento óptimo. Además, deben estar en formatos compatibles como PDF, XML o TXT, ya que estos son los formatos admitidos por los modelos LLM para aplicar eficazmente la técnica RAG. Asegurarse de cumplir con estos requisitos es fundamental para obtener resultados precisos y relevantes durante el entrenamiento del modelo.

- **Chucks**

Se crean los Chucks para dividir los documentos en varias partes más pequeñas, esto logra que el modelo encuentre los fragmentos específicos de información y no necesariamente buscar en todos los documentos completos. Además, los modelos LLM como GPT-3-turbo tiene límites de cantidad de texto para procesar. Si se dividen los textos, se logra asegurar los fragmentos relevantes.

```
from langchain.text_splitter import RecursiveCharacterTextSplitter

text_splitter = RecursiveCharacterTextSplitter(
    chunk_size=1550, ## cada texto va a tener un tamaño de 1500 caracteres
    chunk_overlap=200,
    length_function=len
)

documents = text_splitter.split_documents(ml_papers)
```

- **Embeddings y base de datos vectorial.** Estos son componentes claves, permiten una búsqueda eficiente y precisa.

Embeddings son representación numérica que captura el significado semántico del texto para este proyecto se usó **text-embedding-ada-002** y las bases de datos vectoriales es donde se almacenan y gestionan los Embeddings de manera más eficiente, se usó **Chroma**.

- **Configuración del recuperador de documentos**

```

from langchain.embeddings import OpenAIEmbeddings
from langchain.vectorstores import Chroma

embeddings = OpenAIEmbeddings(model = 'text-embedding-ada-002') ## Embeddings

## creacion de base de datos vectorial usando chroma
vectorstore = Chroma.from_documents(
    documents=documents,
    embedding=embeddings
)

### Configuracion del recuperador de documentos
retriever = vectorstore.as_retriever(
    search_kwargs={"k":3}
)
    
```

- **Modelo de consulta**

Para el modelo de consulta se debe configurar la API de OpenIA para poder usar el modelo LLM que se va a usar, para ello se configuró el modelo de inteligencia artificial Generativa GPT-3.5-turbo

```

from langchain_openai import ChatOpenAI
from langchain.chains import RetrievalQA

chat = ChatOpenAI(
    openai_api_key = OPENAI_API_KEY,
    model_name= 'gpt-3.5-turbo', ## modelo
    temperature= 0.0 ## control de aleatoriedad de las respuestas, es decir no queremos que el modelo sea creativo
)

## combinacion del modelo generativo con el de recuperacion
qa_chain = RetrievalQA.from_chain_type(
    llm= chat,
    chain_type= "stuff", ## se especifica la manera como se debe integrar la informacion recuperada tambien esta map-reduce y refine
    retriever=retriever
)
    
```

7.4 Pruebas

Para llevar a cabo las pruebas básicas, se formula la misma pregunta tanto al modelo base como al modelo entrenado utilizando la técnica RAG. Se realizó un conjunto de preguntas, dividiendo este en; Preguntas Generales, Preguntas Específicas y Preguntas fuera de contexto, logrando identificar que como se comporta el modelo a los diferentes tipos de preguntas.

Preguntas Generales:

Pregunta 1. Modelo base - gpt 3.5 turbo.

```
response = llm_gpt3([
    HumanMessage(content="segun el avance tecnico Semillas biofortificadas que son Las semillasbiofortificadas de maíz y frijol?")
])
print(response)

<ipython-input-6-59a2402f23b4>:1: LangChainDeprecationWarning: The method `BaseChatModel.__call__` was deprecated in langchain-core 0.1.
response = llm_gpt3([
content='Las semillas biofortificadas de maíz y frijol son variedades mejoradas genéticamente que han sido desarrolladas para tener un m
```

Question = ¿Segun el avance tecnico Semillas biofortificadas que son Las semillas biofortificadas de maíz y frijol?

```
response = llm_gpt3([
content='Las semillas biofortificadas de maíz y frijol son variedades mejoradas genéticamente que han sido desarrolladas para tener un mayor contenido de nutrientes esenciales, como vitaminas y minerales. Estas semillas son una herramienta importante en la lucha contra la malnutrición, ya que pueden ayudar a mejorar la dieta de las personas que dependen en gran medida de estos cultivos como fuente de alimento. Las semillas biofortificadas de maíz y frijol pueden contribuir a reducir la deficiencia de nutrientes en poblaciones vulnerables y mejorar la salud y el bienestar de las comunidades agrícolas.'
```

Ampliación de la pregunta y respuesta generada por el modelo

Modelo reentrenado (RAG)

```
query = " segun el avance tecnico Semillas biofortificadas que son Las semillas biofortificadas de maiz y frijol?"
qa_chain.run(query)

<ipython-input-16-157be9be85ad>:2: LangChainDeprecationWarning: The method `Chain.run` was deprecated in langchain 0.1.0 and will be removed in 1
qa_chain.run(query)

'Las semillas biofortificadas de maiz y frijol son variedades mejoradas genéticamente para tener un mayor contenido de nutrientes como zinc y hie
las han sido seleccionadas para garantizar un alto potencial productivo y nutricional en las condiciones específicas de la zona cafetera colombiana'
```

Question = ¿Segun el avance tecnico Semillas biofortificadas que son Las semillas biofortificadas de maiz y frijol?

Content = “Las semillas biofortificadas de maiz y frijol son variedades mejoradas genéticamente para tener un mayor contenido de nutrientes como zinc y hierro. Estas semillas han sido seleccionadas para garantizar un alto potencial productivo y nutricional en las condiciones específicas de la zona cafetera colombiana”.

Ampliación de la pregunta y respuesta modelo reentrenado (RAG)

Pregunta 2. Modelo base gpt - 3.5 turbo.

```
response = llm_gpt3([
    HumanMessage(content="cuales son Las principales tendencias relativas a la producción de cafés sostenibles y la"
        "vinculación de las nuevas generaciones en la producción de café sostenible para el fortalecimiento Del empalme"
        "generacional en el municipio de Aguadas (Caldas)?")
])
print(response)

content='Algunas de las principales tendencias relativas a la producción de cafés sostenibles en el municipio de Aguadas (Caldas) y la vinculación'
```

Question: ¿Cuáles son Las principales tendencias relativas a la producción de cafés sostenibles y la vinculación de las nuevas generaciones en la producción de café sostenible para el fortalecimiento Del empalme generacional en el municipio de Aguadas (Caldas)?

Content = 'Algunas de las principales tendencias relativas a la producción de cafés sostenibles en el municipio de Aguadas (Caldas) y la vinculación de las nuevas generaciones en esta actividad son:\n\n1. Certificaciones de café sostenible: Cada vez más consumidores están buscando cafés que cuenten con certificaciones de sostenibilidad, como Fair Trade, Rainforest Alliance o UTZ. Esto impulsa a los productores a adoptar prácticas sostenibles en sus cultivos. \n\n2. Agricultura orgánica: La demanda de café orgánico está en aumento, lo que lleva a los productores a implementar prácticas agrícolas respetuosas con el medio ambiente, como el uso de abonos naturales y la eliminación de pesticidas químicos.\n\n3. Involucramiento de las nuevas generaciones: Para fortalecer el empalme generacional en la producción de café sostenible, es importante involucrar a las nuevas generaciones en el negocio familiar. Esto puede lograrse a través de programas de capacitación, acceso a tecnologías innovadoras y la promoción de la agricultura sostenible como una actividad atractiva y rentable.\n\n4. Uso de tecnología: La implementación de tecnologías como sistemas de riego eficientes, sensores de humedad del suelo y drones para el monitoreo de cultivos puede ayudar a mejorar la productividad y la sostenibilidad de las fincas cafeteras.\n\n5. Alianzas estratégicas: La colaboración entre productores, cooperativas, empresas y organizaciones gubernamentales puede ser clave para impulsar la producción de café sostenible en Aguadas y garantizar su viabilidad a largo plazo.\n\nEn resumen, las principales tendencias en la producción de café sostenible en Aguadas (Caldas) incluyen la adopción de prácticas sostenibles, el involucramiento de las nuevas generaciones, el uso de tecnología y la creación de alianzas estratégicas

Ampliación de pregunta y respuesta generada por el modelo base

Modelo Reentrenado (RAG)

```
query = (" cuales son Las principales tendencias relativas a la producción de cafés sostenibles y la vinculación de las nuevas"
| "generaciones en la producción de café sostenible para el fortalecimiento Del empalme generacional en el municipio de Aguadas (Caldas)?"
qa_chain.run(query)

'Las principales tendencias relativas a la producción de cafés sostenibles y la vinculación de las nuevas generaciones en la producción de café sostenible para el fortalecimiento del empalme generacional en el municipio de Aguadas (Caldas) incluyen la necesidad de implementar estrategias que fomenten la sostenibilidad ambiental, social y económica en la producción de café. Además, es importante promover la certificación de prácticas sostenibles, el uso de tecnologías amigables con el medio ambiente y la creación de incentivos para atraer a las nuevas generaciones a participar en la caficultura de manera sostenible. Sin embargo, es crucial abordar la falta de interés de los jóvenes en continuar con la tradición cafetera familiar y encontrar formas innovadoras de involucrarlos en el sector.'
```

Question = ¿Cuáles son Las principales tendencias relativas a la producción de cafés sostenibles y la vinculación de las nuevas generaciones en la producción de café sostenible para el fortalecimiento del empalme generacional en el municipio de Aguadas (Caldas)?

Content = "Las principales tendencias relativas a la producción de cafés sostenibles y la vinculación de las nuevas generaciones en la producción de café sostenible para el fortalecimiento del empalme generacional en el municipio de Aguadas (Caldas) incluyen la necesidad de implementar estrategias que fomenten la sostenibilidad ambiental, social y económica en la producción de café. Además, es importante promover la certificación de prácticas sostenibles, el uso de tecnologías amigables con el medio ambiente y la creación de incentivos para atraer a las nuevas generaciones a participar en la caficultura de manera sostenible. Sin embargo, es crucial abordar la falta de interés de los jóvenes en continuar con la tradición cafetera familiar y encontrar formas innovadoras de involucrarlos en el sector."

Ampliación de preguntas y repuestas modelo reentrenado (RAG).

El proceso comienza formulando preguntas generales tanto al modelo base como al modelo reentrenado. Al analizar las respuestas generadas por ambos, se observa que estas son relativamente similares. Esto se debe a que el modelo base puede acceder a la información relevante contenida en sus datos preentrenados, manteniendo una coherencia entre la pregunta y la respuesta. Sin embargo, es importante destacar que las respuestas proporcionadas por el modelo reentrenado mediante la técnica RAG son más puntuales y específicas relacionadas con el texto. Esto demuestra que dichas respuestas se construyen aprovechando la información adicional proporcionada durante el proceso de reentrenamiento.

Preguntas Específicas:

Pregunta 3 Modelo base -gpt-3.5-turbo

```
# Realizar la consulta al modelo utilizando el formato correcto de mensajes
response = llm_gpt3([
    HumanMessage(content="según el avance técnico Cenicafé 547 que se realizó para determinar los niveles de acumulación de zinc de híbrido biofortificado?")
])

<ipython-input-7-6e8c95d1d785>:2: LangChainDeprecationWarning: The method `BaseChatModel.__call__` was deprecated in langchain-core 0.1.7 and will be removed in
response = llm_gpt3([

print(response)

content='El avance técnico Cenicafé 547 determinó que el híbrido biofortificado acumula niveles significativamente más altos de zinc en comparación con otras va
```

Question = "según el avance tecnico cenicafe 547 que se realizó para determinar los niveles de acumulación del zinc biofortificados?"

content='El avance técnico Cenicafé 547 determinó que el híbrido biofortificados acumula niveles significativamente más altos de zinc en comparación con otras variedades de café. Esto es importante porque el zinc es un micronutriente esencial para la salud humana y su presencia en el café puede contribuir a mejorar la nutrición de las personas que consumen esta bebida.'

Ampliación de la respuesta generada por el modelo base

Modelo Reentrenado (RAG)

```
query = " según el avance tecnico cenicafe 547 que se realizo para determinar los niveles de acumulacion de zinc de hibrido biofortificado ?"
qa_chain.run(query)
```

```
'Según el avance técnico Cenicafé 547, se recolectaron aleatoriamente muestras de granos de maíz con 14% de humedad en cuatro estaciones experimentales de Cenicafé para determinar los niveles de acumulación de zinc del híbrido de maíz blanco biofortificado (SGBIOH2). Sin embargo, el documento no proporciona los resultados específicos encontrados en relación con los niveles de acumulación de zinc en el híbrido biofortificado.'
```

Question = "según el avance tecnico cenicafe 547 que se realizó para determinar los niveles de acumulación del zinc biofortificados?"

Content ='Según el avance técnico Cenicafé 547, **se recolectaron aleatoriamente muestras de granos de maíz con 14% de humedad en cuatro estaciones experimentales** de Cenicafé para

Ampliación de la respuesta generada por el modelo (RAG)

Pregunta 4 Modelo base -gpt-3.5-turbo

```
response = llm_gpt3([
    HumanMessage(content="segun el avance tecnico cenicafe 547 cuanto y donde se encontro el nivel de mas"
    "alto de zinc en las muestras recolectadas que contenian 14% de humedas?")
])
print(response)
```

```
<ipython-input-7-9d8b4a344ef7>:1: LangChainDeprecationWarning: The method `BaseChatModel.__call__` was deprecated in langchain-core 0.1.7 and will be removed in :
response = llm_gpt3([
content='Según el avance técnico de Cenicafé 547, el nivel más alto de zinc en las muestras recolectadas que contenían un 14% de humedad se encontró en la región
```

Question = ¿Segun el avance tecnico cenicafe 547 **cuan to y don de se encon tró el nivel de más alto de zinc** en las muestras recolectadas que contenían 14% de húmedas?

Content ='Según el avance técnico de Cenicafé 547, el nivel más alto de zinc en las muestras recolectadas que contenían un 14% de humedad se encontró en la región de Antioquia, con un contenido de 45 ppm de zinc.

Ampliación de la pregunta y respuesta generada por el modelo base

Modelo Reentrenado (RAG)

```
query = "segun el avance tecnico cenicafe 547 cuanto y donde se encontro el nivel de mas alto de zinc en las muestras recolectadas que contenian 14% de humedas?"
qa_chain.run(query)

<ipython-input-18-93f339b0dcde>:2: LangChainDeprecationWarning: The method `Chain.run` was deprecated in langchain 0.1.0 and will be removed in 1.0. Use :meth:`
qa_chain.run(query)
'Según el avance técnico de Cenicafe 547, el nivel más alto de zinc en las muestras recolectadas que contenían 14% de humedad se encontró en la localidad de El R
osario, con un contenido promedio de zinc de 40,18 mg kg-1.'
```

Question = segun el avance tecnico cenicafe 547 cuanto y donde se encontró el nivel de más alto de zinc en las muestras recolectadas que contenían 14% de húmeda?

Content = Según el avance técnico de Cenicafe 547, el nivel más alto de zinc en las muestras recolectadas que contenían 14% de humedad se encontró en la **localidad de El Rosario**, con un **contenido promedio de zinc de 40,18 mg kg-1**.

Ampliación de la pregunta y respuesta generada por el modelo reentrenado

Pregunta 5. Modelo base gpt-3.5. turbo.

```
response = llm_gpt3([
    HumanMessage(content="en que porcentaje y cuales son los oficiones que los estudiantes de las instituciones rurales del municipio de aguadas caldas proyecta
"])
print(response)

<ipython-input-8-00745af95064>:1: LangChainDeprecationWarning: The method `BaseChatModel.__call__` was deprecated in langchain-core 0.1.7 and will be removed in :
response = llm_gpt3([
content="No dispongo de información específica sobre el porcentaje y los oficios que los estudiantes de las instituciones rurales del municipio de Aguadas, Calda:
```

Question = ¿En qué porcentaje y cuáles son los oficios que los estudiantes de las instituciones rurales del municipio de aguadas caldas proyectan como proyecto de vida?

Content='No dispongo de información específica sobre el porcentaje y los oficios que los estudiantes de las instituciones rurales del municipio de Aguadas, Caldas, proyectan como proyecto de vida. Sin embargo, es común que en zonas rurales los estudiantes aspiren a trabajar en actividades relacionadas con la agricultura, ganadería, artesanía, turismo rural, entre otros. También es posible que algunos estudiantes aspiren a estudiar carreras universitarias en áreas como la educación, la salud, la administración, entre otras. Sería necesario realizar un estudio específico en el municipio de Aguadas para obtener datos precisos sobre este tema.'

Ampliación de la pregunta y respuesta generada por el modelo base

Modelo Reentrenado (RAG)

```
query = "en que porcentaje y cuales son los oficiones que los estudiantes de las instituciones rurales del municipio de aguadas caldas proyectan como proyecto de vida"
qa_chain.run(query)

<ipython-input-17-493e89179777>:2: LangChainDeprecationWarning: The method `Chain.run` was deprecated in langchain 0.1.0 and will be removed in 1.0. Use `chain.invoke` instead.
qa_chain.run(query)
'El 60% de los egresados de Instituciones Educativas rurales del municipio de Aguadas, Caldas, se proyectan a emprender en la venta de servicios como barberos, estilistas, mecánicos, y comerciantes. Mientras que el 25% de ellos visualizan sus proyectos de vida en oficios y servicios diferentes a la producción y comercialización de café.'
```

Question = porcentaje y cuáles son las aficiones que los estudiantes de las instituciones rurales del municipio de aguadas caldas proyectan como proyecto de vida?

Content = "El 60% de los egresados de Instituciones Educativas rurales del municipio de Aguadas, Caldas, se proyectan a emprender en la venta de servicios como barberos, estilistas, mecánicos, y comerciantes. Mientras que el 25% de ellos visualizan sus proyectos de vida en oficios y servicios diferentes a la producción y comercialización de café."

Ampliación de la pregunta y respuesta generada por el modelo reentrenado

Para la sección de las preguntas específicas con un total de 3 ejemplos, se formularon preguntas clave orientadas a obtener información puntual contenida en los documentos utilizados durante el reentrenamiento del modelo. Estas preguntas fueron planteadas tanto al modelo base como al modelo reentrenado.

Como resultado, se observó que el modelo base no era capaz de generar respuestas precisas a las preguntas formuladas. En su lugar, proporcionaba un texto relacionado con el tema de la pregunta, pero sin abordar de manera específica lo solicitado. Cabe destacar la pregunta 5, en la cual el modelo base indicó explícitamente que no disponía de la información específica requerida, limitándose a generar una respuesta más general.

Pregunta Fuera de Contexto .

Pregunta 6.

Modelo Reentrenado gpt-3.5 turbo

```
query = " principales características del azúcar?"
qa_chain.run(query)

<ipython-input-16-6026edd44a1e>:2: LangChainDeprecationWarning: The method `Chain.run` was deprecated in langchain
qa_chain.run(query)
'No tengo información específica sobre las principales características del azúcar en el contexto proporcionado.'
```

Question = ¿principales características del azúcar?

Content = **No tengo información específica** sobre las principales características del azúcar en el contexto proporcionado

Pregunta 7.

Modelo Reentrenado GPT-3.5 turbo

```
query = "cuales son las principales actividades la produccion platano?"
qa_chain.run(query)

'Lo siento, pero la información proporcionada no menciona directamente las principales actividades de producción de plátano. te?'
```

Question = ¿Cuáles son las principales actividades la producción del plátano?"

Content = **Lo siento, pero la información proporcionada no menciona directamente** las principales actividades de producción de plátano. ¿Hay algo más en lo que pueda ayudarte?

Al analizar las respuestas generadas por el modelo IAG reentrenado en un contexto diferente al de la información contenida en los documentos, se observa que el modelo se limita a responder que no dispone de la información solicitada o que la pregunta no se encuentra mencionada en los documentos proporcionados.

Considero que este tipo de comportamientos, pueden ayudar a evitar especulación por parte del modelo o generar información incorrecta.

8. Conclusiones

- Para la implementación del modelo RAG es fundamental garantizar la compatibilidad entre las versiones de las librerías y herramientas utilizadas.
- El modelo entrenado con RAG utiliza el contexto proporcionado por los documentos específicos entregados, lo que permite generar respuestas más precisas y detalladas, evitando la generalidad característica del modelo base.
- Se observa que el modelo base tiende a generar respuestas más generales. Esto se debe a que estos modelos, previamente entrenados, generan texto basado en la información contenida en su entrenamiento original. Aunque su respuesta puede ser aproximada, no logra ser tan específica como el modelo reentrenado.
- Es importante tener presente que los datos con los que se desea reentrenar el modelo deben estar formateados con las condiciones necesarias para que el modelo evitar errores al momento de generar las respuestas.
- El modelo IAG reentrenado prioriza la precisión al limitarse a la información disponible en los documentos, evitando especulaciones.

Referencias

- A CIO and CTO technology guide to generative AI | McKinsey.* (n.d.). Retrieved January 15, 2025, from <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/technologies-generational-moment-with-generative-ai-a-cio-and-cto-guide>
- Astridgiraldo/proyecto_Rag: modelo RAG.* (n.d.). Retrieved January 23, 2025, from https://github.com/Astridgiraldo/proyecto_Rag
- Colombia avanza en la adopción de la Inteligencia Artificial Generativa - Fedesoft.* (n.d.). Retrieved January 23, 2025, from https://fedesoft.org/colombia-avanza-en-la-adopcion-de-la-inteligencia-artificial-generativa-el-29-de-las-empresas-estan-en-fase-de-experimentacion-activa-revela-sondeo-de-fedesoft/?utm_source=chatgpt.com
- de la Torre, J. (2023). *REDES GENERATIVAS ADVERSARIAS (GAN) FUNDAMENTOS TEÓRICOS Y APLICACIONES SURVEY.*
- El estado de la IA en 2022 y el balance de media década | McKinsey.* (n.d.). Retrieved January 15, 2025, from https://www.mckinsey.com/featured-insights/destacados/el-estado-de-la-ia-en-2022-y-el-balance-de-media-decada/es?utm_source=chatgpt.com
- Fine-tuning - OpenAI API.* (n.d.). Retrieved January 16, 2025, from <https://platform.openai.com/docs/guides/fine-tuning/common-use-cases>
- Generative AI Data in '25: Importance & 7 Methods.* (n.d.). Retrieved January 15, 2025, from <https://research.aimultiple.com/generative-ai-data/>
- Howard, J., & Ruder, S. (n.d.). *Universal Language Model Fine-tuning for Text Classification.* Retrieved January 22, 2025, from <http://nlp.fast.ai/ulmfit>.
- “Más de 77% de las Pyme en Colombia quieren empezar a usar inteligencia artificial.”* (n.d.). Retrieved January 15, 2025, from <https://www.larepublica.co/empresas/entrevista-con-daniel-verswyvel-gerente-general-de-microsoft-colombia-sobre-la-inteligencia-artificial-en-colombia-3883542>
- Prompt engineering - OpenAI API.* (n.d.). Retrieved January 22, 2025, from <https://platform.openai.com/docs/guides/prompt-engineering>
- ¿Qué es el fine-tuning? | IBM.* (n.d.). Retrieved January 22, 2025, from <https://www.ibm.com/es-es/topics/fine-tuning>

¿Qué es RAG?: explicación de la IA de generación aumentada por recuperación, AWS. (n.d.).

Retrieved January 16, 2025, from <https://aws.amazon.com/es/what-is/retrieval-augmented-generation/>

Szép, M., Rueckert, D., Von Eisenhart-Rothe, R., & Hinterwimmer, F. (2024). *A Practical Guide to Fine-tuning Language Models with Limited Data*. <https://huggingface.co/models>

Vista de Análisis de las estrategias de producción de café sostenible para el fortalecimiento Del empalme generacional en el municipio de Aguadas (Caldas). (n.d.). Retrieved January 23, 2025, from

<https://revistas.ucp.edu.co/index.php/entrecienciaeingenieria/article/view/2900/2713>

Vista de Maíz y frijol biofortificados intercalados con café | Avances Técnicos Cenicafé. (n.d.).

Retrieved January 22, 2025, from https://publicaciones.cenicafe.org/index.php/avances_tecnicos/article/view/274/334

What is a Variational Autoencoder? | IBM. (n.d.). Retrieved January 15, 2025, from <https://www.ibm.com/think/topics/variational-autoencoder#Overview>

What is AI Training Data? Explore AI Training Datasets & Providers | Datarade. (n.d.). Retrieved January 15, 2025, from <https://datarade.ai/data-categories/ai-ml-training-data>

What is LangChain? - LangChain Explained - AWS. (n.d.). Retrieved January 16, 2025, from <https://aws.amazon.com/what-is/langchain/>

What is LLM? - Large Language Models Explained - AWS. (n.d.). Retrieved January 15, 2025, from <https://aws.amazon.com/what-is/language-model/>