Modelos de lenguaje para difusión del conocimiento agrícola.

Language Models for Agricultural Knowledge difussion.

Brayan Rusell Figueroa Orantes (1).
Estudiante de Licenciatura, Tecnológico Nacional de México/I. T. de Tuxtla Gutiérrez.
L20270766@tuxtla.tecnm.mx.

Diego Zapata García (2). Estudiante de Licenciatura, Tecnológico Nacional de México/I. T. de Tuxtla Gutiérrez, L20270826@tuxtla.tecnm.mx.

Germán Ríos Toledo* (3). Tecnológico Nacional de México/I. T. de Tuxtla Gutiérrez, german.rt@tuxtla.tecnm.mx.

Aída Guillermina Cossio Martínez (4). Tecnológico Nacional de México/ I. T. de Tuxtla Gutiérrez, aida.cm@ittg.edu.mx.

Néstor Antonio Morales Navarro (5). Tecnológico Nacional de México/ I. T. de Tuxtla Gutiérrez, nestor.mn@tuxtla.tecnm.mx.

Karen Daniela Ramón Cristancho (6). Estudiante de Maestría, Tecnológico Nacional de México/ I. T. de Tuxtla Gutiérrez, M24271381@tuxtla.tecnm.mx.

*corresponding author.

Artículo recibido en octubre 30, 2025; aceptado en diciembre 05, 2025.

Resumen.

En nuestro país, la agricultura enfrenta desafíos como la baja productividad, la degradación del suelo y la contaminación de agua. Por otro lado, existe la falta de difusión del conocimiento agrícola en la población en general. Para abordar este problema, en este trabajo se propone la implementación de un Modelo de Lenguaje Grande y Generación Aumentada por Recuperación. El prototipo utiliza archivos de texto con datos agrícolas generales sobre la planta de arroz, plagas y enfermedades comunes, así como recomendaciones de riego y fertilizantes. La información de estos archivos se almacena en una base de datos vectorial. Cuando el usuario realiza consultas en lenguaje natural, el recuperador de información realiza una búsqueda por similitud en la base de datos vectorial. La información recuperada se inyecta al Modelo de Lenguaje para generar respuestas confiables. La implementación de este modelo representa una estrategia innovadora para superar las limitaciones en la difusión del conocimiento agrícola, promoviendo prácticas más eficientes y sostenibles.

Palabras claves: Base de datos vectorial, búsqueda por similitud, generación aumentada por recuperación, modelo de lenguaje grande.

Abstract.

In our country, agriculture faces challenges such as low productivity, soil degradation, and water pollution. On the other hand, there is a lack of dissemination of agricultural knowledge among the general population. To address this problem, this paper proposes the implementation of a Large Language Model and Retrieval-Augmented Generation.

The prototype uses text files with general agricultural data on rice plants, common pests and diseases, as well as irrigation and fertilizer recommendations. The information in these files is stored in a vector database. When the user makes queries in natural language, the information retriever performs a similarity search in the vector database. The retrieved information is injected into the Language Model to generate reliable responses. The implementation of this model represents an innovative strategy to overcome the limitations in the dissemination of agricultural knowledge, promoting more efficient and sustainable practices.

Keywords: Large Language Model, retrieval-augmented generation, similarity search, vector database.

1. Introducción.

La economía de un país se estructura tradicionalmente en tres grandes sectores: primario, secundario y terciario. El sector primario comprende actividades esenciales como la agricultura, la ganadería, la pesca y la explotación forestal, las cuales representan la base de los recursos que sostienen al resto de la cadena productiva. El sector secundario transforma esas materias primas en bienes industriales, mientras que el sector terciario integra los servicios que dinamizan el comercio, el transporte, la educación y otras actividades fundamentales para el crecimiento económico. Dentro de esta estructura, la agricultura juega un papel fundamental al generar alimentos, materias primas y empleo, contribuyendo de manera directa al Producto Interno Bruto y al fortalecimiento de la economía nacional. Como lo plantea el Comité Estatal de Información y finanzas (2024), el estado de Chiapas se encuentra ubicado en el sureste de México, es un estado que cuenta con abundantes recursos naturales. De acuerdo con cifras reportadas por el Comité Estatal de Información Estadística y Geográfica de Chiapas (CEIEG) (Comité Estatal de Información y finanzas. (2024), de los 32 estados que conforman la república, Chiapas ocupó el 14º lugar nacional por el valor de la producción agrícola en 2024, con una participación de 2.71% del total nacional, lo que representó un monto de 23,207.2 millones de pesos. En la **Tabla 1** se muestra en valor de la producción de los 12 cultivos predominantes en Chiapas. Se observa la importancia que tiene la producción de maíz, la caña de azúcar y el café para la economía del estado.

Tabla 1. Valor de la Producción Agrícola por Cultivo Principal de Chiapas.

Cultivos Principales	**Valor de la producción (millones de pesos)**
Maíz grano	7,195,558
Caña de Azúcar	2,940,035
Café Cereza	2,150,205
Resto de los cultivos	2,023,416
Plátano	1,708,921
Mango	1,492,232
Palma africana	1,243,952
Frijol	1,024,950
Pastos y praderas	948,605
Papaya	824,528
Tomate rojo	797,064
Aguacate	491,401
Cacao	366,362
TOTAL	**23,207,227**

Por otro lado, la **Figura 1** muestra la proporción de producción de los mismos cultivos respecto a las 8,853,275 millones de toneladas reportadas en el informe de CEIG para el año 2024. La caña de azúcar es el cultivo que más se produce, seguido de pastos y pradera (para la ganadería) y el maíz.

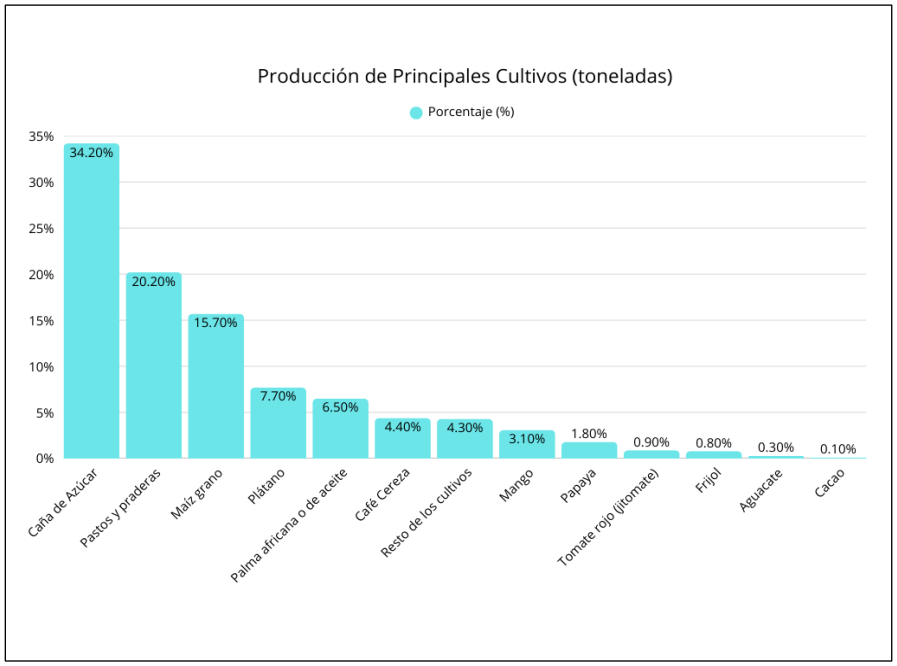

Figura 1. Proporción de la Producción Agrícola Anual de Chiapas según el Informe CEIG 2024.

El gobierno federal tiene como eje principal de su política, impulsar la agricultura para lograr la autosuficiencia en la producción de cultivos más importantes del país a pequeña y gran a escala. El objetivo es factible de alcanzar para los productores a gran escala, ya que cuentan con recursos económicos que les permiten tecnificar el proceso de producción desde la siembra hasta la cosecha. Además, tienen acceso a información actualizada de las más recientes investigaciones y de las tendencias del mercado. Sin embargo, a pequeña escala el panorama es distinto. Si bien muchas personas podrían cultivar sus propios cultivos, esto no ocurre frecuentemente porque para hacerlo de forma eficiente y sostenible, se requieren años de experiencia en la siembra, el cuidado y la prevención de plagas, aunado a habilidades técnicas y prácticas que no están al alcance inmediato de quienes no han trabajado previamente en la agricultura. Más aún, dado que los pequeños productores se basan en gran medida en el conocimiento empírico, no se genera la experiencia necesaria difundir ese conocimiento de forma consistente y ponerla al alcance de más personas.

En ese sentido, los avances en Inteligencia Artificial, especialmente en los Modelos Generativos de Lenguaje, se han convertido en herramientas valiosas para poner al alcance de millones de personas el conocimiento disponible sobre prácticamente cualquier tema de interés. Estos modelos permiten acceder de forma rápida y sencilla a información clara y útil, incluida aquella relacionada con la agricultura, desde técnicas de siembra hasta prácticas de cuidado y manejo de cultivos. Cabe mencionar que el solo hecho del acceso al conocimiento no resuelve el problema, pero permite la auto capacitación de las personas interesadas en ciertos cultivos.

Un modelo generativo de lenguaje es un sistema de IA entrenado con grandes corpus de texto que aprende patrones lingüísticos y produce texto nuevo, coherente y contextualmente relevante, lo que permite su aplicación en diversos ámbitos como la agricultura (Patel *et al*., 2025; Hagos *et al*., 2024). La **Tabla 2** muestra algunos de los Modelos de Lenguaje más utilizados en la actualidad.

Tabla 2. Principales Modelos de Lenguaje Grandes.

Modelo de Lenguaje Grande	URL oficial / ubicación
GPT-4 / GPT-5 (OpenAI)	https://openai.com
Claude 3 / Claude 4 (Anthropic)	https://www.anthropic.com
Gemini (Google DeepMind)	https://deepmind.google
LLaMA 3 (Meta AI)	https://ai.meta.com/llama
Grok (xAI, Elon Musk)	https://x.ai

2. Métodos.

En el panorama actual de la inteligencia artificial corporativa, la mera capacidad de generar texto ya no es suficiente. Las organizaciones requieren sistemas que "razonen" basándose en sus propios datos, políticas y archivos históricos. No obstante, los Modelos de Lenguaje Grande (LLM) presentan un inconveniente: su conocimiento del mundo queda congelado hasta el momento de su entrenamiento. Para superar esta barrera, la industria ha adoptado la arquitectura de Generación Aumentada por Recuperación (**RAG**, **R**etrieval **A**ugmented **G**eneration).

Siguiendo los parámetros de la empresa NVIDIA (2023) y la **Figura 2** el flujo de trabajo de un sistema con RAG de cada etapa, desde el preprocesamiento hasta la recuperación semántica, es vital para garantizar la calidad del resultado final. En el ámbito de la Inteligencia Artificial Generativa, la precisión de los modelos depende intrínsecamente de la calidad y estructura de los datos que consumen. A continuación, se describen cada una de las etapas.

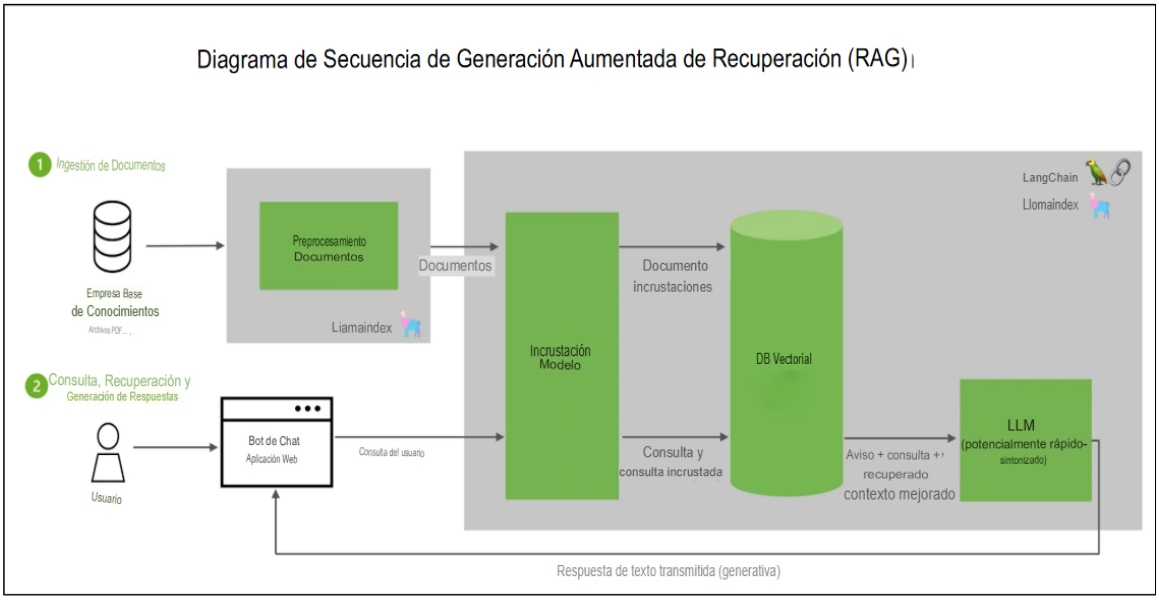

Figura 2. Diagrama de secuencia de recuperación y generación aumentada (RAG) de NVIDIA.

Preprocesamiento de documentos:

Antes de que los datos puedan ser utilizados, como indica Lewis *et al.* (2020) es necesario convertirlos en un formato legible para los modelos de inteligencia artificial. Esto implica tomar documentos en formatos variados como PDF, Word o presentaciones y extraer el texto contenido en ellos, asegurándose de que el texto sea consistente y sea legible para la máquina. Esto implica el uso de reconocimiento de caracteres. Así mismo, también existe texto dentro de las

tablas y los gráficos. Este proceso asegura que los datos mantengan su significado y no se pierda la relación entre cifras y conceptos.

Extracción de metadatos (opcional):

Esta etapa, según Jenna Pederson (2025) es el equivalente a colocar una etiqueta clara en el lomo de cada libro. No se trata solo del contenido, sino de quién lo dijo y cuándo. Para optimizar la recuperación futura de información, se recomienda extraer datos descriptivos del documento, conocidos como metadatos, tales como el Título del documento, el Autor o fuente, el Número de páginas y la URL de origen si el material proviene de la web. Esta información adicional funciona como un índice avanzado, permitiendo que el sistema no solo sepa *qué* dice el documento, sino también *de dónde* proviene, lo cual es vital para la indexación y la priorización de fuentes confiables.

Fragmentación de documentos:

Como explican Schwaber-Cohen y Patel (2025), el *chunking* es una técnica que consiste en dividir información grande o compleja en partes más pequeñas y manejables llamadas *chunks*. Su objetivo principal es facilitar la comprensión, el aprendizaje y el procesamiento de la información, ya sea para personas (por ejemplo, al estudiar o leer) o para sistemas de inteligencia artificial (como al dividir textos largos para analizarlos mejor).

Al crear chunks, es importante cuidar algunos parámetros clave. Uno de ellos es el tamaño del *chunk*, que debe ser lo suficientemente grande para mantener una idea completa, pero no tan extenso como para volver a ser difícil de procesar. Otro parámetro relevante es el solapamiento entre *chunks*, que consiste en repetir una pequeña parte del contenido entre fragmentos consecutivos para no perder contexto.

Por ello, los documentos largos deben ser divididos en fragmentos más pequeños y manejables. El objetivo es asegurar que cada pedazo sea lo suficientemente pequeño para ser procesado por el modelo, pero a la vez, que mantenga la coherencia semántica interna, es decir, que la idea principal no se corte a mitad. Al final, esto mejora notablemente la precisión al recuperar información, permitiendo al sistema recuperar solo la sección específica necesaria sin tener que procesar el documento entero.

Generación de embeddings:

Cada fragmento de texto se convierte en un "vector numérico" o *embedding* mediante modelos de representación semántica. Este proceso traduce el texto a un espacio matemático vectorial, lo que permite representar el contenido del fragmento en dicho espacio. Los textos con significados similares se agruparán muy juntos, lo que permite comparar fragmentos basándose en su significado real y no solo en la coincidencia de palabras clave. Esto es esencial para facilitar la búsqueda rápida y eficiente en bases de datos vectoriales.

Indexación de embeddings:

Los *embeddings* generados se almacenan en una base de datos vectorial. Este paso permite: que los miles de vectores se guardarlos en una estructura que permita encontrarlos instantáneamente. Este almacenamiento estructurado permite recuperar fragmentos de manera eficiente durante la búsqueda, ya que se pueden aplicar técnicas matemáticas de "búsqueda de similitud" (como la distancia coseno o Euclidiana) para encontrar los datos más relevantes. Además, este paso es crucial para mantener actualizada la información disponible en el sistema de recuperación.

Consulta del usuario:

Cuando una persona realiza una pregunta, su consulta pasa por el mismo proceso de transformación: se convierte en un *embedding* para capturar la intención y el contexto de lo que se pregunta. Este nuevo vector de consulta se lanza al espacio vectorial de la base de datos para poder comparar la pregunta con los *embeddings* almacenados y buscar los fragmentos de documentos más similares a la pregunta.

Recuperación de información relevante:

El sistema ejecuta una búsqueda en la base de datos vectorial seleccionando los fragmentos que están "más cerca" matemáticamente de la pregunta del usuario. Esto se logra principalmente mediante la búsqueda de similitud semántica, que encuentra los fragmentos con mayor relación lógica con la pregunta. Adicionalmente, se puede aplicar un filtrado adicional basado en los metadatos extraídos anteriormente para afinar aún más la precisión de los resultados.

Generación de respuesta:

Solo una vez que se han recuperado los fragmentos verificados, el proceso finaliza. Los fragmentos de información altamente relevantes que se recuperaron son pasados a un Modelo de Lenguaje Grande (LLM). El LLM actúa como un redactor inteligente: fusiona múltiples fragmentos si es necesario, y redacta una respuesta fluida, clara y coherente para el usuario, basándose únicamente en la información encontrada. Por último, un sistema completo debe proporcionar referencias a los documentos de origen cuando sea relevante, dando validez a la respuesta generada.

3. Desarrollo.

Para llevar a cabo la experimentación, se utilizó el entorno *Colab* debido a que cuenta con recursos de hardware de altas prestaciones, principalmente la GPU y TPU requeridos para ejecutar modelos de lenguaje con fluidez.

No obstante, se presentaron limitaciones en cuanto a los límites de uso y los tiempos de espera, lo cual impactó en la fluidez del proceso. La experimentación incluyó una recopilación de información de diversas fuentes especializadas en el ámbito agrícola como Climate-Smart Agriculture in Chiapas, México (Zavariz-Romero. 2014), Cultivo de Arroz: Técnicas y Consejos para una Siembra Exitosa (Verdejo, 2024), y Cultivo de arroz: como se realiza, plagas y enfermedades como dicta el Agrotendencia (2025). Sin embargo, el modelo de lenguaje utilizado para la implementación del proyecto Implementación de un modelo LLM Local para su uso en cultivos del Estado de Chiapas son una combinacion de Modelo de lenguaje (LLM) DeepSeek-R1:8B integrado con LangChain usando OllamaLLM para respuestas en RAG y modelos de embeddings (vectorización de texto) que son all-mpnet-base-v2 que maneja un mejor rendimento en la similitudo de semanticas junto a BAAI/bge-base-en que optimiza las tareas de retrieval que se representa en biblioteca de sentence-transformers, el cual estaba restringido al idioma inglés. Para superar esta limitación, se empleó Google Translate para modificar y adaptar los datos, lo que permitió generar preguntas relevantes basadas en la información obtenida.

Finalmente, estas preguntas fueron utilizadas para evaluar el rendimiento del modelo a lo largo de las diferentes etapas del proceso experimental. A continuación, en las **Tablas 3**, **4** y **5**, se presentan las preguntas generadas para cada conjunto de datos.

Tabla 3. Preguntas generadas para Arroz-Agrotendencia.

No.	Arroz-Agrotendencia.txt
1	What is the historical origin of rice cultivation according to ancient Chinese manuscripts?
2	How was rice cultivation introduced to Latin America, and who contributed to its adaptation?
3	What does the scientific term Oryza sativa mean, and what is its origin?
4	Why is rice considered the second most consumed cereal worldwide?
5	What notable nutritional properties does rice have compared to other cereals?
6	How does rice’s adaptability relate to different climatic conditions?
7	Which countries were the top rice producers in 2016 according to FAO data?
8	How does the concentration of rice exporters affect international prices?
9	What are the three agronomic phases of rice’s growth cycle?
10	How do nodal and crown roots contribute to the physiology of rice plants?

Las respuestas a cada pregunta se encuentran en el siguiente enlace:
https://drive.google.com/file/d/1ILrGhELs64chwKLu3Wf_MDWJHsYraOEq/view?usp=sharing.

Tabla 4. Preguntas generadas para Arroz-mundoagricultura.

No.	Arroz-mundoagricultura.txt
1	What are the most common methods for planting rice, and how do they differ?
2	What are the optimal climatic and soil conditions for rice cultivation?
3	What factors should be considered in water management for successful rice farming?
4	What nutritional differences exist between polished rice and palay rice?
5	What steps are involved in harvesting rice, and why must it not be delayed?
6	How long does rice take to mature depending on its variety and cultivation method?
7	What regions in Mexico are the main rice producers, and how is their production distributed?
8	How does soil type help prevent issues like waterlogging in rice cultivation?
9	What practices are recommended for pest and disease control in rice farming?
10	How does temperature affect the development and germination of rice?

Las respuestas a cada pregunta se encuentran en el siguiente enlace:
https://drive.google.com/file/d/1y_GYg4L6RUmF03y_xJbhIYuN92NV7B9e/view?usp=sharing.

Tabla 5. Preguntas generadas para CSA-en-Chiapas-Mexico.

No.	CSA-en-Chiapas-Mexico.txt
1	What is Climate-Smart Agriculture (CSA) and what are its three main pillars?
2	How does agroforestry contribute to CSA in Chiapas?
3	What percentage of Chiapas' Gross Domestic Product (GDP) is contributed by agriculture?
4	What are the main crops produced in Chiapas?
5	What are the key climate risks affecting agriculture in Chiapas?
6	What percentage of greenhouse gas (GHG) emissions in Chiapas comes from agricultural activities?
7	Why is minimum tillage in corn production considered a CSA practice?
8	What are some CSA strategies that can help farmers in Chiapas cope with extreme weather?
9	How does land fragmentation impact agricultural productivity in Chiapas?
10	What role do institutions and policies play in promoting CSA in Chiapas?

Las respuestas a cada pregunta se encuentran en el siguiente enlace:
https://drive.google.com/file/d/10jeiNyGyCUZqc5P0uvXX1CFhov5n9vYM/view?usp=sharing.

Resultados de los experimentos.

En este apartado se presentan y analizan los resultados obtenidos tras la evaluación exhaustiva del sistema propuesto. El objetivo principal es medir la eficacia de la implementación de la técnica de Generación Aumentada por Recuperación (RAG) frente a un modelo de lenguaje (LLM) estándar utilizando su conocimiento base.

La experimentación se llevó a cabo utilizando tres conjuntos de datos específicos: Agrotendencia, Mundoagricultura y CSA en Chiapas, México. Para garantizar una validación integral, el análisis se ha estructurado en tres etapas secuenciales que permiten aislar y evaluar cada componente del flujo de trabajo. Para cuantificar el rendimiento y la precisión semántica en cada etapa, no se utilizaron comparaciones textuales simples, sino que se emplearon dos modelos de *embeddings* (representación vectorial) considerados estados del arte en el procesamiento de lenguaje natural. Estos modelos actúan como "jueces" para determinar qué tan similar es el significado de la respuesta generada con respecto a la respuesta esperada:

MPNET (Masked and Permuted Pre-training): Es un modelo robusto conocido por su alta precisión en tareas de similitud semántica. Evalúa qué tan bien captura el modelo la estructura y el significado profundo de las oraciones, ofreciendo una medida fiable de la coherencia del texto.

BGE (BAAI General Embedding): Desarrollado por la *Beijing Academy of Artificial Intelligence*, es uno de los modelos de recuperación y embedding más avanzados actualmente. Se utiliza en este experimento para contrastar los resultados de MPNET y ofrecer una segunda validación de alta sensibilidad sobre la relevancia de la información.

Etapa 1: Preguntas directamente al modelo.

La evaluación busca medir la capacidad del modelo para generar respuestas basadas únicamente en su entrenamiento previo, sin utilizar datasets externos o procesos de recuperación como RAG. Esto permite establecer una línea base de conocimiento y determinar si el modelo tiene información preexistente sobre los temas o contextos asociados a cada conjunto de datos.

Como se muestra en la **Tablas 6**, **7** y **8**, los puntajes de los modelos MPNET Y BGE representan la **similitud del coseno**. Un porcentaje más alto indica que la respuesta generada es semánticamente más cercana a la respuesta ideal o al fragmento de referencia.

Tabla 6. Agrotendencia || resultados obtenidos del modelo.

Pregunta	MPNET	BGE	Tiempo (S)
1	74.17	86.60	58.04
2	80.34	91.83	18.20
3	77.70	91.53	22.58
4	64.10	85.24	27.95
5	83.04	89.93	30.88
6	86.71	89.48	32.64
7	84.21	91.87	24.11
8	83.53	89.27	34.95
9	87.14	92.93	18.61
10	82.36	90.23	23.53

Tabla 7. Mundoagricultura || resultados obtenidos del modelo.

Pregunta	Puntaje MPNET	Puntaje d BGE %	Tiempo (S)
1	79.14	89.15	44.41
2	77.99	91.32	37.91
3	75.61	88.77	24.55
4	88.11	90.06	36.92
5	78.41	87.95	32.26
6	79.12	85.99	33.02
7	82.84	90.45	26.64
8	84.44	91.61	30.40
9	77.28	88.52	36.11
10	79.12	89.43	28.27

Tabla 8. CSA en Chiapas México || resultados obtenidos del modelo.

Pregunta	Puntaje MPNET	Puntaje BGE	Tiempo (S)
1	81.48	92.25	23.12
2	68.94	89.74	30.372
3	71.12	94.86	22.623
4	79.97	91.42	28.821
5	82.15	91.90	42.602
6	76.22	92.61	29.762
7	74.27	88.52	26.924
8	78.99	89.82	35.562
9	80.27	90.77	23.92
10	53.48	87.39	38.791

En la **Figuras 3** y **4**, se muestra el rendimiento general del modelo basado en las puntuaciones de similitud obtenidas durante las consultas. El gráfico de pastel refleja cómo se distribuyen las respuestas en 6 categorías de similitud: Muy malo (0-50 (%)), Malo (50-60 (%)), Regular (60-70 (%)), Bueno (70-80 (%)), Muy bueno (80-90 (%)), Excelente (90-100 (%)).

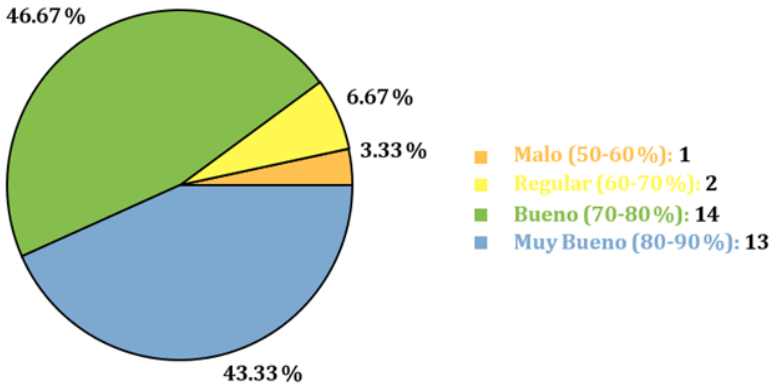

Figura 3. Distribución de puntaje de similitud MPNET.

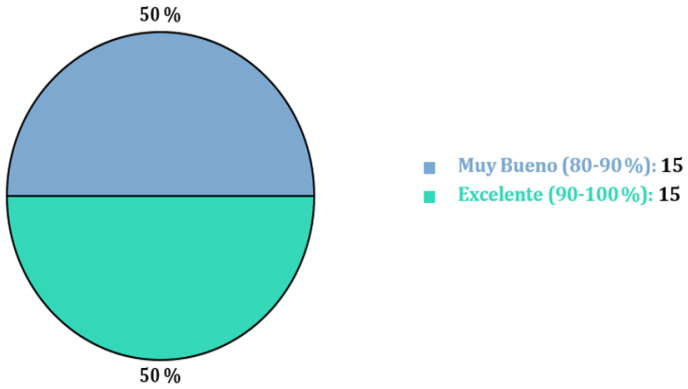

Figura 4. Distribución de puntaje de similitud BGE.

Estos gráficos confirman la importancia de integrar sistemas como RAG y bases de datos especializadas (como ChromaDB o PineCone) para mejorar la precisión y relevancia de las respuestas del modelo.

Etapa 2: Preguntas Chromadb.

El propósito de esta sección es verificar que la base de datos vectorial, creada mediante ChromaDB, haya sido configurada correctamente y que contenga todos los documentos relevantes. En la **Tablas 9**, **10** y **11**, el proceso asegura que la información proporcionada al modelo esté disponible de manera íntegra y organizada, lo cual es fundamental para el éxito de las consultas realizadas durante las etapas posteriores. Además, esta verificación permite confirmar que todos los documentos han sido incorporados y que no existen datos faltantes que puedan afectar la recuperación de información.

Tabla 9. Agrotendencia || resultados obtenidos de Chromadb.

Pregunta	Fragmento A		Fragmento B		Fragmento C	
	BGE	MPNET	BGE	MPNET	BGE	MPNET
1	88.06	62.45	86.03	58.57	86.03	58.57
2	88.86	75.13	86.13	78.08	86.13	78.08
3	87.52	42.12	83.33	33.12	76.76	40.08
4	89.92	70.90	85.40	71.12	86.28	73.36
5	88.77	73.56	85.39	74.85	86.47	71.19
6	84.57	55.16	82.54	64.68	83.52	67.78
7	85.48	66.74	82.38	56.98	83.82	58.57
8	87.48	72.81	81.25	57.29	81.25	57.29
9	93.92	86.83	83.53	62.79	82.45	65.32
10	87.67	67.40	81.45	55.50	81.45	55.50

Tabla 10. Mundoagricultura || resultados obtenidos de Chromadb.

Pregunta	Fragmento A		Fragmento B		Fragmento C	
	BGE	MPNET	BGE	MPNET	BGE	MPNET
1	85.03	74.05	83.90	74.77	83.90	74.77
2	87.75	72.09	81.95	62.86	84.59	70.08
3	84.20	56.61	82.85	52.78	83.39	57.14
4	87.83	67.36	80.44	45.56	79.92	57.70
5	86.56	70.76	87.76	76.09	88.05	74.63
6	88.08	70.26	81.93	69.13	84.02	78.74
7	84.81	69.62	83.25	60.67	83.25	60.67
8	86.42	70.77	82.05	53.36	84.08	60.79
9	82.89	53.28	82.32	53.10	79.94	53.12
10	85.34	64.99	82.00	55.82	84.18	68.00

Tabla 11. CSA || resultados obtenidos de Chromadb.

Pregunta	Fragmento A		Fragmento B		Fragmento C	
	BGE	MPNET	BGE	MPNET	BGE	MPNET
1	88.43	72.21	88.11	66.91	88.11	66.91
2	89.31	69.56	86.56	68.45	84.52	67.96
3	89.49	67.22	84.06	57.59	81.69	56.35
4	87.06	74.41	82.85	56.45	80.44	53.88
5	90.80	70.48	87.49	62.26	85.96	69.14
6	93.23	77.69	86.04	61.04	83.31	61.96
7	87.07	61.45	83.40	48.86	82.90	57.38
8	88.09	76.68	86.46	64.88	84.66	68.44
9	88.98	69.53	85.22	49.76	84.00	56.30
10	89.11	68.51	86.62	55.23	84.65	55.89

En las **Figuras 5** y **6** podemos observar la distribución del rendimiento general de los fragmentos basado en las puntuaciones de similitud recuperados de la base de datos vectorial ChromaDB.

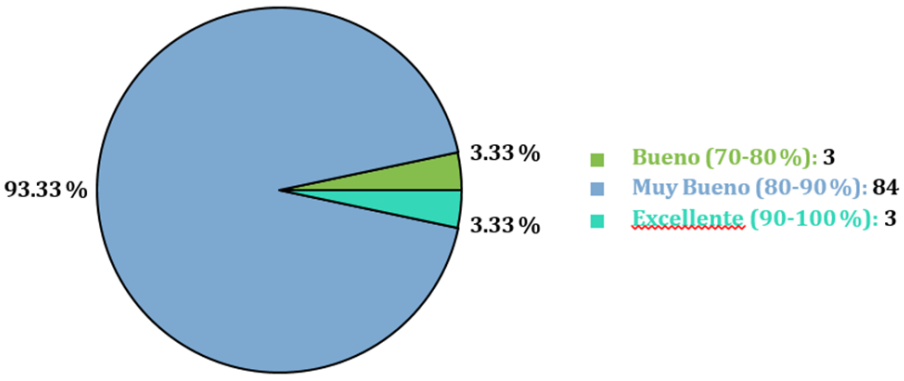

Figura 5. Distribución de puntaje de similitud BGE || Fragmentos.

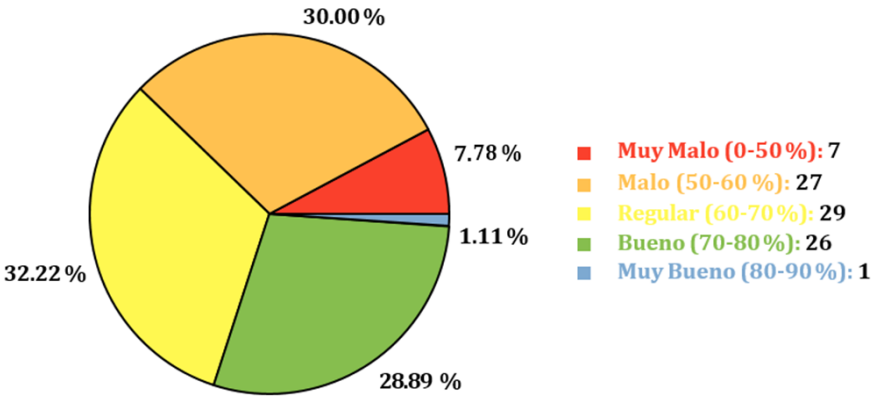

Figura 6. Distribución de puntaje de similitud MPNET || Fragmentos.

Se identifican áreas en las que la recuperación de fragmentos de las bases de datos vectoriales no resulta completamente relevante. Los gráficos presentados destacan la importancia de validar el conocimiento proporcionado al modelo LLM, asegurando que el análisis de la información suministrada sea preciso y alineado con la consulta realizada. Este proceso es fundamental para optimizar la capacidad del modelo de generar respuestas altamente pertinentes y enriquecedoras para el usuario.

Etapa 3: Preguntas después de la implementación de la técnica RAG.

En esta última etapa en la **Tablas 12**, **13** y **14**, se evalúa el impacto de la implementación de la técnica RAG (Generación Aumentada por Recuperación) en el modelo LLM, utilizando la base de datos ChromaDB.

RAG fue implementado para mejorar la capacidad del modelo de generar respuestas más precisas y relevantes, al permitirle recuperar información de una base de datos externa en lugar de depender exclusivamente de su entrenamiento previo. Esto permite al modelo acceder a datos adicionales almacenados en los documentos, lo que mejora significativamente la calidad, relevancia y el enfoque de las respuestas, especialmente en áreas donde el modelo original podría carecer de conocimiento directo.

Tabla 12. Agrotendencia || resultados obtenidos RAG.

Pregunta	Respuesta		Fragmento A		Fragmento B		Fragmento C	
	BGE	MPNET	BGE	MPNET	BGE	MPNET	BGE	MPNET
1	92.88	80.36	88.06	62.45	83.61	57.82	82.09	59.17
2	91.37	82.08	87.82	82.28	88.86	75.13	87.66	72.57
3	93.90	83.58	87.52	42.12	79.72	47.76	78.68	44.09
4	89.64	78.38	89.92	70.90	85.55	62.67	84.16	60.58
5	90.78	82.09	88.06	76.44	88.77	73.56	82.16	50.16
6	88.71	78.20	84.46	64.22	83.16	62.58	82.31	62.00
7	91.95	82.93	85.48	66.74	84.60	62.41	82.09	54.99
8	90.42	79.65	87.48	72.81	79.03	51.54	77.90	50.32
9	94.85	87.86	93.92	86.83	85.32	67.99	82.95	62.12
10	91.92	78.37	87.67	67.40	81.28	52.33	79.48	51.21

Tabla 13. Mundoagricultura || resultados obtenidos RAG.

Pregunta	Respuesta		Fragmento A		Fragmento B		Fragmento C	
	BGE	MPNET	BGE	MPNET	BGE	MPNET	BGE	MPNET
1	90.21	83.33	85.03	74.05	82.26	68.48	83.53	64.58
2	89.25	81.39	86.21	74.27	87.75	72.09	82.52	61.59
3	87.86	79.42	83.38	59.93	84.20	56.61	81.88	56.47
4	91.63	83.11	87.83	67.36	82.62	62.12	79.66	52.19
5	89.81	81.92	86.56	70.76	85.27	64.99	85.51	60.67
6	94.50	78.13	88.08	70.26	84.02	66.05	82.44	66.37
7	91.81	81.13	84.81	69.62	83.83	61.00	83.19	58.53
8	92.28	81.82	86.42	70.77	84.72	63.98	83.10	64.65
9	91.93	77.89	82.71	56.31	81.99	56.89	80.55	55.68
10	91.69	80.97	85.34	64.99	82.72	60.35	81.42	61.46

Tabla 14. CSA || resultados obtenidos RAG.

Pregunta	Respuesta		Fragmento A		Fragmento B		Fragmento C	
	BGE	MPNET	BGE	MPNET	BGE	MPNET	BGE	MPNET
1	92.33	86.02	88.43	72.21	87.20	73.39	85.68	73.62
2	89.26	64.58	89.07	73.44	88.49	72.73	88.10	72.98
3	93.38	76.43	89.49	67.22	87.75	63.65	84.60	63.60
4	92.36	78.85	87.06	74.41	84.85	67.09	83.13	65.69
5	92.54	80.53	90.80	70.48	87.27	72.50	88.24	70.78
6	93.96	82.89	93.23	77.69	86.88	65.93	87.01	63.46
7	89.51	68.59	87.07	61.45	84.90	62.63	83.54	61.54
8	88.96	81.94	88.09	76.68	87.63	73.78	88.05	71.11
9	91.93	83.30	86.61	73.90	88.98	69.53	86.79	68.81
10	90.46	66.41	89.11	68.51	88.38	66.72	87.41	67.39

En las siguientes **Figuras 7 y 8**, se observa la distribución de la similitud entre la consulta y la respuesta (Query-Response) después de la implementación de RAG. La similitud fue calculada utilizando la similitud del coseno, que mide la cercanía entre dos vectores (en este caso, las consultas y las respuestas) basándose en las palabras comunes que contienen, sin tener en cuenta el significado contextual o semántico profundo.

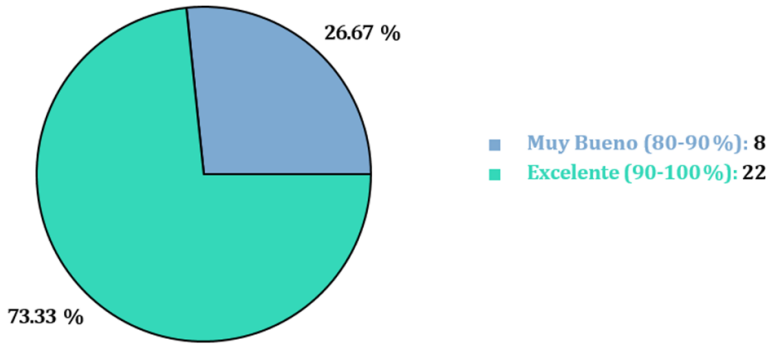

Figura 7. Distribución de puntaje de similitud BGE || RAG.

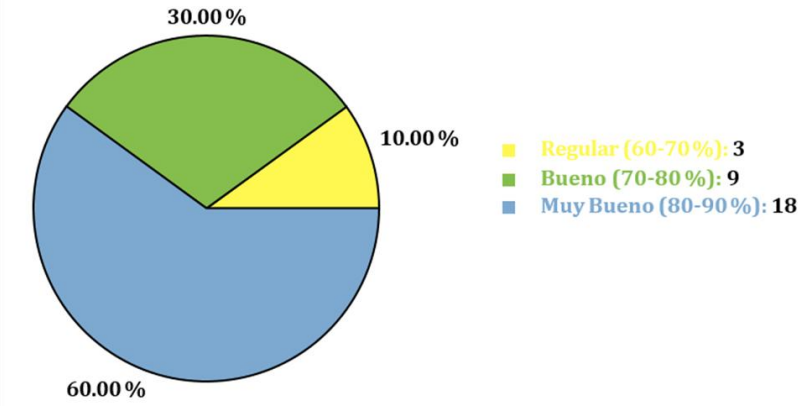

Figura 8. Distribución de puntaje de similitud MPNET || RAG.

Conclusiones.

El proyecto de implementación de un Modelo de Lenguaje Grande (LLM) para cultivos en Chiapas logró configurar el modelo Llama 2 7b, el cual es capaz de responder a preguntas relacionadas con la agricultura. Este avance representa un paso significativo para abordar el déficit de conocimiento agrícola, no solo para los productores, sino también para la población en general.

Durante el desarrollo de este proyecto, se mostró la relevancia de la técnica de Generación Aumentada por Recuperación. Dada su versatilidad, esta técnica permite proporcionar información al modelo sin la necesidad de entrenarlo continuamente, lo que mejora su eficiencia y aplicabilidad en escenarios donde se requiere acceso a datos específicos sin recurrir a un reentrenamiento constante. La Generación Aumentada por Recuperación destaca por su capacidad para integrar información actualizada y específica. Esto permite una mayor eficiencia computacional y reduce los costos asociados al entrenamiento tradicional. Además, facilita la adaptación del sistema a distintos dominios mediante la incorporación de nuevas fuentes de conocimiento. En conjunto, esta técnica mejora la precisión y confiabilidad de las respuestas en contextos especializados. La Generación Aumentada por Recuperación permite a los Modelos de Lenguaje Grande proporcionar respuestas más completas y precisas, favoreciendo la difusión de conocimiento agrícola accesible a los productores y la comunidad en general. La inclusión de información externa amplió la perspectiva del sistema, mejorando su aplicabilidad y utilidad en el contexto local, y contribuyendo a la mejora del entendimiento y la adopción de mejores prácticas agrícolas en la región.

Créditos.

Los autores agradecen al Tecnológico Nacional de México por el financiamiento del proyecto a través de la convocatoria de Proyectos de Investigación Científica, Desarrollo Tecnológico e Innovación 2025.

Referencias bibliográficas.

Agrotendencia. (2025). Cultivo de arroz: conoce cómo se realiza y sus plagas. Agrotendencia.tv. https://agrotendencia.tv/agricultura/cultivos/cereales/el-cultivo-de-arroz/

Comité Estatal de Información y Finanzas. (2024). Cuaderno de información agrícola 2024. https://www.ceieg.chiapas.gob.mx/storage/posts/productos/CIGECH/Cuaderno_Agricultura_2024.pdf

Hagos, D. H., Battle, R., & Rawat, D. B. (2024). Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *IEEE transactions on artificial intelligence*.

Jenna Pederson. (2025). Pinecone. https://www.pinecone.io/learn/retrieval-augmented-generation/

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/2005.11401

NVIDIA. (2023). RAG 101: Retrieval-Augmented Generation Pipeline. NVIDIA Technical Blog. https://developer.nvidia.com/blog/rag-101-demystifying-retrieval-augmented-generation-pipelines/

Patel, R., et al. (2025). Large Language Models in 2024–2025: Trends, Techniques and Challenges. Journal of Emerging Management Studies.

Schwaber-Cohen Roie & Arjun Patel. (2025). Chunking Strategies for LLM Applications. https://www.pinecone.io/learn/chunking-strategies/?utm_source=chatgpt.com

Verdejo, E. (2024, May 15). Cultivo de arroz: Consejos y para una siembra exitosa. Mundo Agricultura. https://mundoagricultura.com/cultivos/cultivo-de-arroz-como-se-da-el-arroz/

Zavariz-Romero. (2014). Climate-smart agriculture in Chiapas, Mexico. CGIAR Research Program on Climate Change, Agriculture and Food Security https://cgspace.cgiar.org/server/api/core/bitstreams/3095a84a-033d-47c0-991e-a0ebba40855a/content

Información de los autores.

Brayan Rusell Figueroa Orantes, Ingeniero en Sistemas Computacionales egresado del Instituto Tecnológico de Tuxtla Gutiérrez, curso la especialidad en tecnología web y móvil aplicada al comercio electrónico, realizo su tesis profesional con el proyecto denominado “Implementación de un modelo LLM Local para su uso en cultivos del Estado de Chiapas”.

Diego Zapata García, Ingeniero en Sistemas Computacionales egresado del Instituto Tecnológico de Tuxtla Gutiérrez, curso la especialidad en tecnología web, realizo su tesis profesional con el proyecto denominado “Implementación de un modelo LLM Local para su uso en cultivos del Estado de Chiapas”.

Germán Ríos Toledo, obtuvo el grado de Doctor en Ciencias de la Computación en 2019 por el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) en Cuernavaca, Morelos, México. Actualmente, es profesor de tiempo completo en el Departamento de Computación del Tecnológico Nacional de México/ I.T. de Tuxtla Gutiérrez, Chiapas), imparte materias en la licenciatura en Ingeniería en Sistemas Computacionales y en la Maestría en Ciencias en Ingeniería Mecatrónica. Su área de especialización es el Procesamiento del Lenguaje Natural.

Aída Guillermina Cossío Martínez es Maestra en Ciencias en Administración por el Instituto Tecnológico de Tuxtla Gutiérrez en 2002. Es profesora de tiempo completo del área de Ingeniería en Sistemas Computacionales del Instituto Tecnológico de Tuxtla Gutiérrez, desde 1994. Se especializa en la formulación y evaluación de proyectos, así como el emprendimiento y desarrollo de planes de negocio, perfil deseable desde 2015 y trabaja en la línea de investigación Sistemas de Información, actualmente miembro del Cuerpo Académico Tecnología de Información para el Desarrollo de Ventajas Competitivas.

Néstor Antonio Morales Navarro. Doctor en Desarrollo Tecnológico de la Universidad Descartes. Maestro en Ciencias en Ingeniería Mecatrónica. Ingeniero en Sistemas Computacionales del Instituto Tecnológico de Tuxtla Gutierrez. Miembro del Sistema Nacional de Investigadores nivel Candidato del Conahcyt. Miembro del Sistema Estatal de Investigadores del estado de Chiapas como Investigador Científico Honorifico. Tiene el reconocimiento a Profesor de Tiempo Completo con Perfil Deseable. Sus intereses de investigación incluyen visión por computadora, aprendizaje profundo y automatización en sistemas mecatrónicos.

Karen Daniela Ramón Cristancho. Ingeniera Electrónica egresada de la Universidad de Pamplona (Colombia). Actualmente cursa la Maestría en Ciencias en Ingeniería Mecatrónica en el Tecnológico Nacional de México, Instituto Tecnológico de Tuxtla Gutiérrez (Chiapas), donde desarrolla trabajo académico en el área de sistemas mecatrónicos inteligentes aplicados a la agricultura de precisión.