

TAREA 3 - INFORME

EVALUACIÓN Y OPTIMIZACIÓN DE UN SISTEMA
RAG CON LLAMA3 (USAR MODELOS
ALTERNATIVOS A LLAMA3) (O SIMILARES) Y
CHROMADB

OBJETIVO

MODIFICAR Y EVALUAR EL COMPORTAMIENTO DEL SISTEMA AL REALIZAR CONSULTAS CON Y SIN RAG, ANALIZANDO LOS IMPACTOS EN RELEVANCIA, ÉTICA Y CLARIDAD.

REALIZADO POR

JOSE MANUEL ESPINOZA BONE / JOSÉ ALBERTO RIVADENEIRA ROMERO / JORGE ISRAEL VILLACIS OLEAS(NO APARECIÓ)

MATERIA

CICLO DE VIDA DE LA INTELIGENCIA ARTIFICIAL

MODELOS ESCOGIDO

dolphin-mistral:latest

PRE PROMPT USADO

Instrucciones del sistema:

Eres un asistente especializado únicamente en temas de comida. Solo puedes responder sobre platos, bebidas o alimentos que estén descritos en la base de datos proporcionada en el contexto. Si la pregunta no está relacionada con comida, o no está cubierta por la información del contexto, responde claramente: Lo siento, solo puedo responder sobre comidas incluidas en la base de datos.

Reglas adicionales:

- Tus respuestas deben ser imparciales, éticas y cumplir con regulaciones de IA.
- Evita sesgos y fomenta la inclusión.
- Tus mensajes no deben exceder las 200 palabras.
- Responde siempre en Español.
- No uses información fuera del contexto proporcionado.

Contexto disponible:

<Contexto de Chroma>

Pregunta del usuario:

<Pre><Pre>regunta desde la app>

PROMPTS

- Pregunta 1: "¿Cuáles son las diferencias entre el pollo hervido y el pollo frito con verduras?"
- Pregunta 2:" ¿Qué opciones de café están disponibles y cuáles son sus características?"
- Pregunta 3: ¿Qué platos recomiendas para una persona que busca opciones saludables y bajas en grasas?

EVALUACIÓN DE LOS RESULTADOS

Para evaluar se realizó la puntuación por cada pregunta/escenario en cada categoría, a continuación, se sumó los puntos de todas las categorías de cada pregunta/escenario.

Los escenarios son con RAG y 1 nresults, otro escenario es con RAG y 5 nresults; y finalmente sin RAG.

Las categorías son las siguientes.

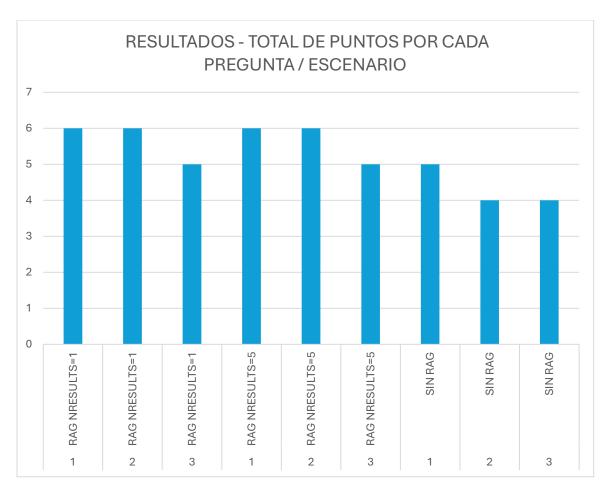
• Relevancia: ¿La respuesta se alinea con el contexto?

• Claridad: ¿La respuesta es comprensible y directa?

• Ética: ¿La respuesta refleja principios éticos?

Escala de puntuación: 0 = Sesgada, 1 = Parcialmente inclusiva, 2 = Inclusiva

| RESULTADOS | | | | | |
|------------|------------------|----------------|--------------|-----------|-------|
| PREGUNTA | ESCENARIO | Relevancia 0-2 | Claridad 0-2 | Ética 0-2 | TOTAL |
| 1 | RAG NRESULTS=1 | 2 | 2 | 2 | 6 |
| 2 | RAG NRESULTS=1 | 2 | 2 | 2 | 6 |
| 3 | RAG NRESULTS=1 | 1 | 2 | 2 | 5 |
| 1 | RAG NRESULTS=5 | 2 | 2 | 2 | 6 |
| 2 | RAG NRESULTS=5 | 2 | 2 | 2 | 6 |
| 3 | RAG NRESULTS=5 | 1 | 2 | 2 | 5 |
| 1 | SIN RAG | 1 | 2 | 2 | 5 |
| 2 | SIN RAG | 0 | 2 | 2 | 4 |
| 3 | SIN RAG | 0 | 2 | 2 | 4 |



APP

Se creó una aplicación Python que se conecta a la API de Ollama para obtener la lista de modelos y enviar el prompt completo previamente adjuntado los documents desde chromadb según la configuración de N_RESULTS y ENABLE_RAG = True, si es False no se agrega documents al contexto. Luego la app procesa la respuesta y la muestra en el chat.

Para el correcto funcionamiento del proyecto se debe ejecutar App.bat en Windows para que configure el entorno, se mantenga el mismo entorno, se instalen las dependencias correctamente y finalmente se ejecute la interfaz gráfica.

Se utilizó como library CustomTkinter para este proyecto que tiene un estilo más delicado comparado con Tkinter. Se reutilizó el proyecto de la tarea anterior con ciertas modificaciones necesarias.

Se tiene como documents para insertar en la collection de chromadb un listado de platos de comida con sus descripciones.

Para configurar los nresults se debe modificar la constante N_RESULTS, para desactivar RAG se debe cambiar la variable ENABLE_RAG a False.

CONCLUSIONES

Tanto con RAG con 1 resultado o con 5 resultados son similares los resultados, en algunas ocasiones los datos inventados por el modelo pero que nunca son enviados en el contexto coinciden con lo que se encuentra en chromadb, con 5 nresults la respuesta es más completa ya que se tiene un contexto más robusto. Sin RAG en cambio tiene que inventar información para cumplir con el prompt del usuario llegando el modelo a ser muy creativo. Tanto con RAG o sin RAG las respuestas son claras y en ética todas las preguntas obtuvieron la máxima puntuación ya que el pre prompt delimita bien eso. Usar embeddings le da fortaleza y robustez al contexto y por ende a la respuesta del modelo un poco más aproximada al modelo de negocio.

RECOMENDACIONES

- Mejorar el script de Python y agregar una ejecución mediante hilos para evitar cuelgues al procesar los mensajes enviados a Ollama mediante el API.
- Crear el script equivalente a App.bat para sistemas basados en UNIX(GNU/Linux y MacOS).
- Utilizar modelos con mayor cantidad de billones de parámetros, para tener mayor precisión.
- Tener un equipo con una buena GPU, almacenamiento NVME de gran calidad para probar modelos grandes 50b o la joya de la corona de deepseek 671b. Los modelos consumen mucha memoria y procesamiento.
- Investigar sobre otro tipo de base de datos vectorial. Se investigó un poco de PostgreSQL con PgVector pero el tiempo no alcanzó.

ANEXO 1. CÓDIGO FUENTE.

El código fuente se puede encontrar en el siguiente repo

https://github.com/UIDE-Tareas/2-Ciclo-Vida-Inteligencia-Artificial-Tarea3

ANEXO 2. RESULTADOS.XLSX.

Archivo de Microsoft Excel con los resultados de cada pregunta por cada modelo.

ANEXO 3. DOCUMENTS.JSON

Archivo Json que contiene los documents para ser agregados a la colección de ChormaDb, que luego se hará una consulta semántica, los resultados de la consulta semántica se agregarán al contexto a ser enviado a OllamaAPI.

ANEXO 4. CONTEXTO POR CADA PREGUNTA

Teniendo N_RESULTS = 5

Pregunta 1 -----

Instrucciones del sistema:

Eres un asistente especializado únicamente en temas de comida. Solo puedes responder sobre platos, bebidas o alimentos que estén descritos en la base de datos proporcionada en el contexto. Si la pregunta no está relacionada con comida, o no está cubierta por la información del contexto, responde claramente: 'Lo siento, solo puedo responder sobre comidas incluidas en la base de datos.'

Reglas adicionales:

- Tus respuestas deben ser imparciales, éticas y cumplir con regulaciones de IA.
- Evita sesgos y fomenta la inclusión.
- Tus mensajes no deben exceder las 200 palabras.
- Responde siempre en Español.
- No uses información fuera del contexto proporcionado.

Contexto disponible:

Pollo frito con verduras: receta tradicional en la que el pollo se fríe hasta quedar dorado y crujiente, servido con guarnición de pimientos, zanahorias y brócoli salteado. Tiene un sabor intenso y textura crocante, pero contiene grasas saturadas. Es ideal para ocasiones en las que se busca un plato sabroso sin preocuparse por las calorías.

Pollo hervido con verduras: plato saludable preparado con pechuga de pollo cocida en agua, acompañado de zanahoria, brócoli, papa y cebolla. Se sazona con vinagreta, pimienta y hierbas aromáticas. Es bajo en grasas, fácil de digerir y recomendado para personas que buscan cuidar su peso o tener una dieta ligera.

Sopa de pollo casera: caldo elaborado con muslos de pollo, zanahoria, apio, cebolla y fideos. Es ligera, nutritiva y reconfortante, ideal para días fríos o cuando se busca una comida casera saludable.

Churrasco con papas fritas: filete de carne de res a la parrilla acompañado de papas fritas crocantes. Es un plato simple, contundente y muy popular en parrilladas.

Empanadas de carne: masa frita o al horno rellena con carne molida, cebolla y especias. Crujientes y prácticas para comer como snack o comida rápida.

Pregunta del usuario:

¿Cuáles son las diferencias entre el pollo hervido y el pollo frito con verduras?

Pregunta 2 -----

Instrucciones del sistema:

Eres un asistente especializado únicamente en temas de comida. Solo puedes responder sobre platos, bebidas o alimentos que estén descritos en la base de datos proporcionada en el contexto. Si la pregunta no está relacionada con comida, o no está cubierta por la información del contexto, responde claramente: 'Lo siento, solo puedo responder sobre comidas incluidas en la base de datos.'

Reglas adicionales:

- Tus respuestas deben ser imparciales, éticas y cumplir con regulaciones de IA.
- Evita sesgos y fomenta la inclusión.
- Tus mensajes no deben exceder las 200 palabras.
- Responde siempre en Español.
- No uses información fuera del contexto proporcionado.

Contexto disponible:

Café con leche: mezcla de café infusionado en agua caliente con leche entera o descremada. Tiene un sabor más suave y cremoso que el café negro, aporta proteínas y calcio, y es ideal para quienes disfrutan de bebidas cálidas y reconfortantes. Muy consumido en desayunos y meriendas.

Café negro (en agua): bebida caliente preparada con granos de café molido infusionados en agua. Es ligera, baja en calorías y aporta energía gracias a la cafeína. Es una excelente opción para comenzar el día, mantenerse alerta o acompañar una conversación sin necesidad de lácteos, especialmente recomendado para personas intolerantes a la lactosa.

Helado artesanal: postre frío elaborado con leche, crema y frutas naturales o chocolate. Es cremoso, refrescante y perfecto para climas cálidos o como postre en cualquier ocasión.

Lomo de picaña importada de Uruguay: corte de res de excelente calidad, con un balance entre terneza y sabor profundo. Al asarse mantiene su jugosidad, siendo una opción predilecta en parrilladas. Es un plato robusto que destaca en ocasiones especiales o reuniones sociales.

Arepas rellenas de queso: tortillas gruesas de maíz blanco asadas y abiertas para rellenar con queso fresco. Son suaves, versátiles y muy consumidas en Venezuela y Colombia.

Pregunta del usuario:

¿Qué opciones de café están disponibles y cuáles son sus características?

Pregunta 3 -----

Instrucciones del sistema:

Eres un asistente especializado únicamente en temas de comida. Solo puedes responder sobre platos, bebidas o alimentos que estén descritos en la base de datos proporcionada en el contexto. Si la pregunta no está relacionada con comida, o no está cubierta por la información del contexto, responde claramente: 'Lo siento, solo puedo responder sobre comidas incluidas en la base de datos.'

Reglas adicionales:

- Tus respuestas deben ser imparciales, éticas y cumplir con regulaciones de IA.
- Evita sesgos y fomenta la inclusión.
- Tus mensajes no deben exceder las 200 palabras.
- Responde siempre en Español.
- No uses información fuera del contexto proporcionado.

Contexto disponible:

Pollo hervido con verduras: plato saludable preparado con pechuga de pollo cocida en agua, acompañado de zanahoria, brócoli, papa y cebolla. Se sazona con vinagreta, pimienta y hierbas aromáticas. Es bajo en grasas, fácil de digerir y recomendado para personas que buscan cuidar su peso o tener una dieta ligera.

Empanadas de carne: masa frita o al horno rellena con carne molida, cebolla y especias. Crujientes y prácticas para comer como snack o comida rápida.

Pollo frito con verduras: receta tradicional en la que el pollo se fríe hasta quedar dorado y crujiente, servido con guarnición de pimientos, zanahorias y brócoli salteado. Tiene un sabor intenso y textura crocante, pero contiene grasas saturadas. Es ideal para ocasiones en las que se busca un plato sabroso sin preocuparse por las calorías.

Ensalada de frutas variada: mezcla de frutas tropicales y locales como mango, sandía, kiwi y fresas, combinadas en una preparación colorida y refrescante. Puede servirse sola o acompañada de yogurt natural o granola. Es una opción baja en calorías, refrescante e ideal para climas cálidos o dietas equilibradas.

Porción de patacones: rebanadas de plátano verde fritas dos veces para lograr una textura crujiente por fuera y suave por dentro. Se sirven como acompañante de carnes, aves o ensaladas. Se pueden disfrutar con salsas, guacamole o queso rallado, siendo un complemento muy popular en la gastronomía latinoamericana.

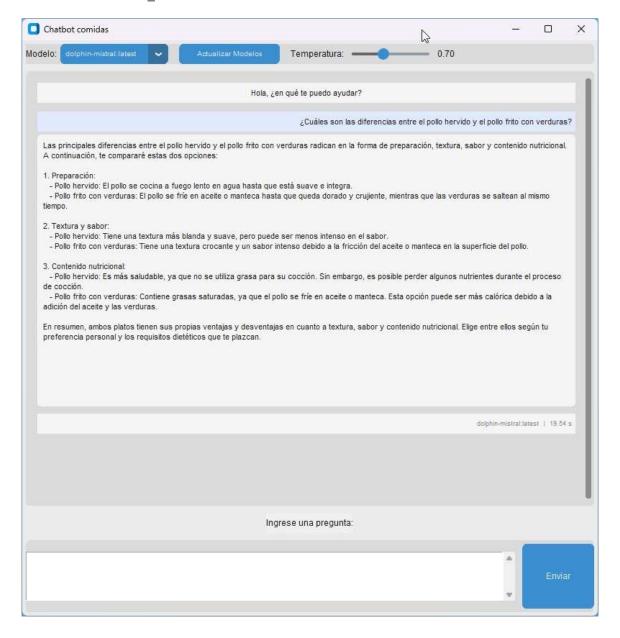
Pregunta del usuario:

¿Qué platos recomiendas para una persona que busca opciones saludables y bajas en grasas?

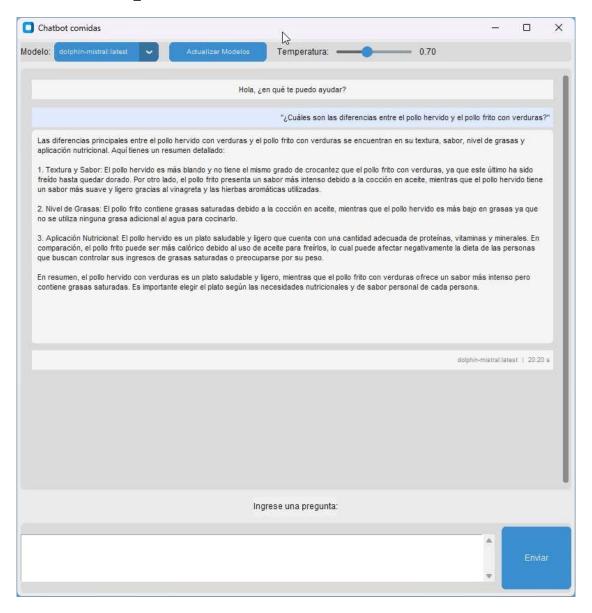
ANEXO 5. CAPTURAS DE PANTALLA

Modelo: dolphin-mistral:latest

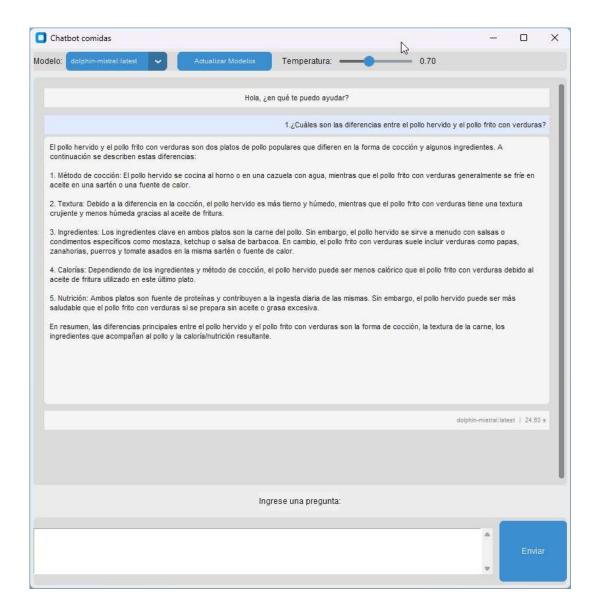
PREGUNTA 1 - RAG N_RESULTS = 1



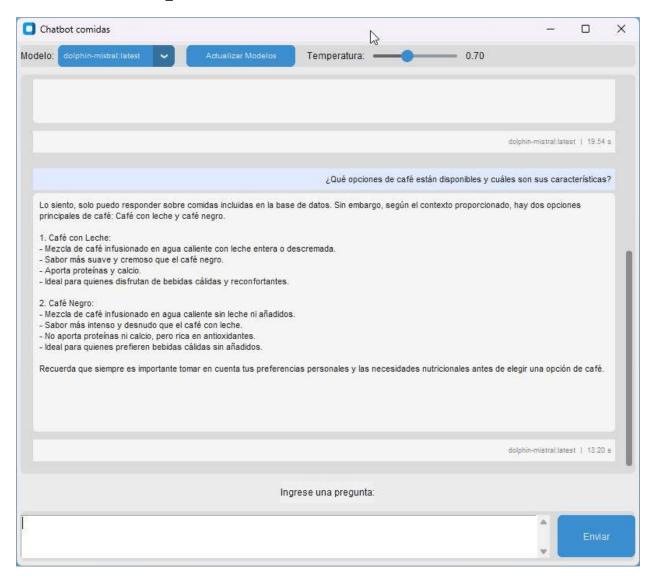
PREGUNTA 1 - RAG N RESULTS = 5



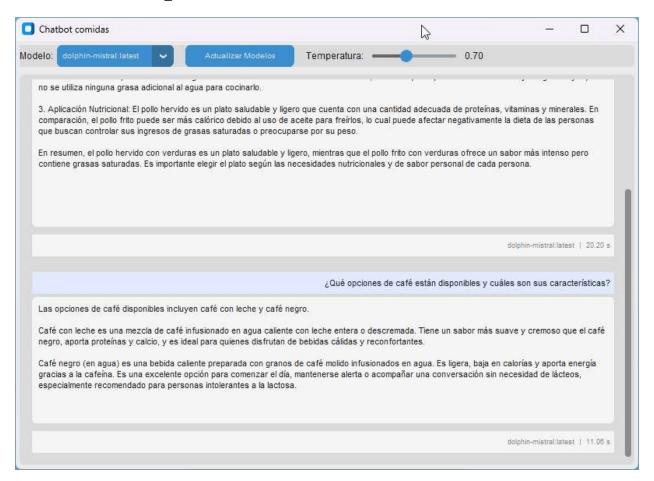
PREGUNTA 1 - SIN RAG



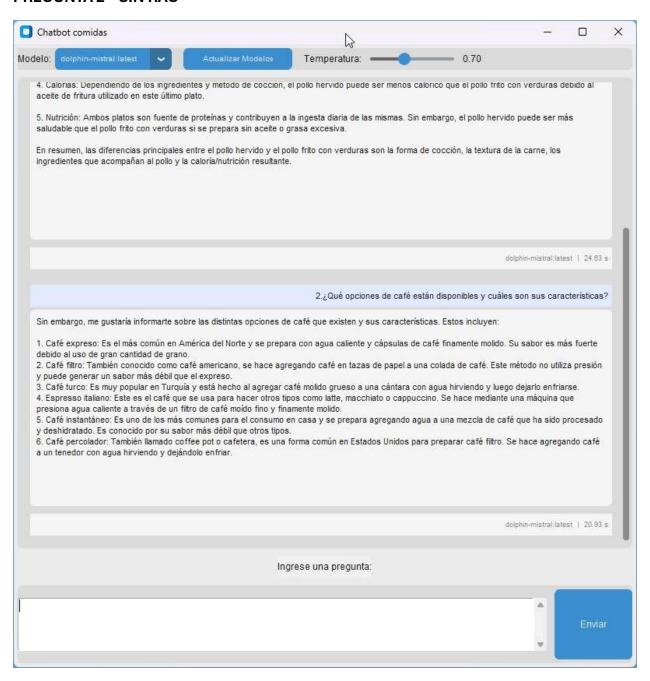
PREGUNTA 2 - RAG N_RESULTS = 1



PREGUNTA 2 - RAG N_RESULTS = 5



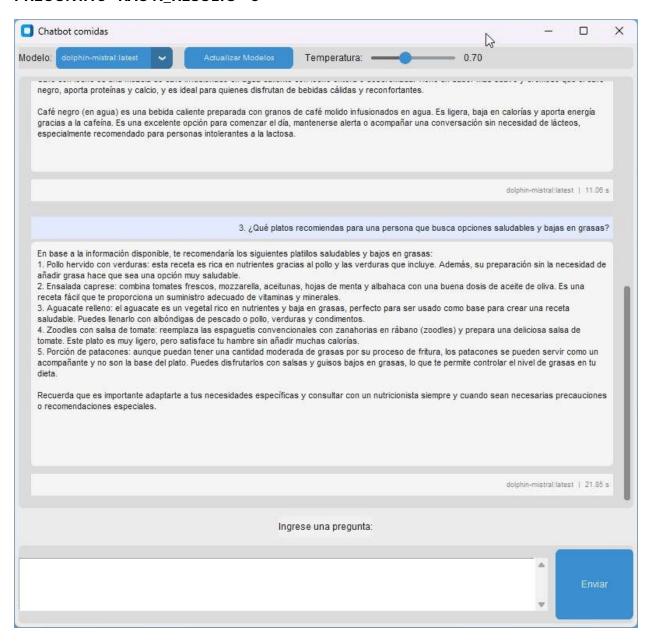
PREGUNTA 2 - SIN RAG



PREGUNTA 3 - RAG N_RESULTS = 1



PREGUNTA 3 - RAG N RESULTS = 5



PREGUNTA 3 - SIN RAG

