Taller 1 - Colgate/Palmolive Chatbot

Universidad Autónoma de Occidente

Facultad de Ingeniería y Ciencias Básicas

Curso: Técnicas Avanzadas de IA Aplicadas a Modelos de Lenguaje

Profesor: Jan Polanco Velasco

Estudiantes: Soren Acevedo - 22500566
Juan Jose Bonilla - 22502052
Yan Carlos Cuaran Imbacuan - 22502591
Nicolas Lozano Mazuera - 22500565

1. Planteamiento de Solución

Descripción general de la empresa

Colgate-Palmolive es una corporación multinacional con presencia en más de 200 países, dedicada principalmente a la producción y comercialización de productos de cuidado oral, personal, del hogar y nutrición animal.

En Colombia, la compañía ha consolidado su presencia como una de las marcas más reconocidas en el sector de bienes de consumo, gracias a sus líneas emblemáticas Colgate (higiene oral) y Palmolive (cuidado corporal y capilar).

La organización se distingue por su compromiso con la innovación, la sostenibilidad y el bienestar social, pilares que le han permitido mantener un liderazgo constante en el mercado global.

Su presencia digital refleja una estrategia comunicativa centrada en la transparencia, la educación del consumidor y la difusión de sus políticas de responsabilidad social y ambiental (ESG), aspectos que fueron fundamentales para construir la base de conocimiento del asistente Q&A.

Arquitectura general del sistema Q&A

Con el objetivo de materializar la propuesta del asistente virtual, se diseñó una arquitectura modular que permite procesar, limpiar y aprovechar la información pública de Colgate-Palmolive para generar respuestas automatizadas mediante modelos de lenguaje.

El sistema se organiza en un pipeline secuencial, compuesto por distintas etapas que garantizan la trazabilidad y la coherencia del flujo de datos, desde la adquisición inicial hasta la interacción final con el usuario.

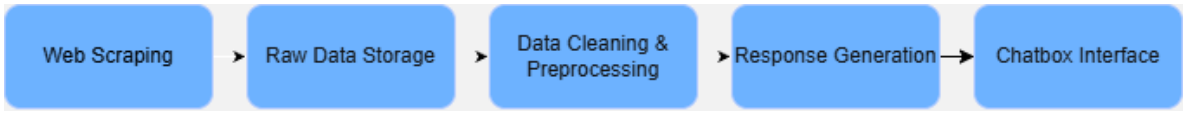

Figura 1. Pipeline etapa inicial Asistente Multimodelo.

El proceso inicia con la etapa de Web Scraping, encargada de recolectar información textual y estructurada de las fuentes oficiales de la empresa.

Los datos capturados se almacenan de forma temporal en una capa de Raw Data Storage, que actúa como repositorio intermedio.

Posteriormente, en la fase de Data Cleaning & Preprocessing, se depura y normaliza la información para garantizar su calidad semántica.

La etapa de Response Generation utiliza este contenido como contexto de referencia para que el modelo de lenguaje genere respuestas precisas y coherentes.

Finalmente, los resultados son desplegados a través de una interfaz de chatbot desarrollada en Streamlit, que permite la interacción directa con el usuario final.

Fuentes de información y estrategia de datos

Para la creación del núcleo de conocimiento semántico, se seleccionaron fuentes públicas, oficiales y actualizadas que reflejan fielmente la identidad, los valores y los servicios de Colgate-Palmolive.

Cada fuente fue inspeccionada manualmente y procesada mediante técnicas de Scraping para garantizar la integridad del texto y la coherencia semántica.

Categoría	**URL**	**Propósito dentro del sistema Q&A**
Historia	colgatepalmolive.com/en-us/who-we-are/history	Responder sobre los orígenes, trayectoria y evolución global de la empresa.
Quiénes somos	colgatepalmolive.com.co/who-we-are	Ofrecer contexto sobre misión, visión, valores y operaciones en Colombia.
Sostenibilidad	colgatepalmolive.com/en-us/impact/sustainability	Describir políticas ambientales, compromisos sociales y metas de impacto sostenible.
Productos Colgate	colgate.com/es-co/products	Responder preguntas sobre categorías y características de productos orales disponibles en el país.
Productos Palmolive	palmolive.co/productos	Brindar información sobre productos de cuidado personal y sus principales beneficios.
Contacto	colgatepalmolive.com.co/contact-us	Proporcionar medios de comunicación, atención al cliente y enlaces de contacto.
YouTube oficial	https://www.youtube.com/colgatecolombia	Obtener información corporativa reciente, logros, campañas y cultura organizacional.

Tabla 1. Fuentes de información a scrapear.

Criterios de selección

- Priorizar fuentes oficiales o verificadas directamente de la empresa.
- Considerar textos en inglés y español para enriquecer la base bilingüe del modelo.
- Incluir solo información vigente, estructurada y de utilidad directa para el usuario final.
- Garantizar el cumplimiento de las políticas de uso de datos (revisión de archivos robots.txt y Términos del sitio).

Alcance del sistema Q&A

El sistema de Preguntas y Respuestas (Q&A) se diseñó como un componente central del asistente virtual de Colgate-Palmolive, orientado a ofrecer información confiable, clara y contextualizada sobre la empresa y sus productos. Su objetivo principal es facilitar la interacción entre los usuarios y la marca, resolviendo inquietudes comunes de forma automatizada y accesible desde una interfaz sencilla.

A través del conocimiento extraído de las fuentes oficiales, el sistema busca cubrir un amplio espectro de temas que reflejan la identidad y las operaciones de la empresa. Entre ellos se incluyen aspectos como la historia de la empresa, la misión, la visión, iniciativas de sostenibilidad y responsabilidad social, relacionadas con el impacto ambiental y las metas de desarrollo sostenible; así como información sobre productos, categorías, beneficios, composición y disponibilidad general en el mercado colombiano.

Además, el sistema integra datos sobre canales de atención y contacto, permitiendo orientar al usuario hacia los medios oficiales de comunicación con la empresa. También incorpora información sobre la presencia digital de Colgate-Palmolive en redes profesionales y sociales.

Si bien en esta etapa inicial el sistema se centra en responder preguntas generales y en la recuperación de información textual, su diseño contempla la futura integración de módulos adicionales que permitirán manejar consultas más específicas, como precios, disponibilidad por punto de venta o atención personalizada al cliente. Esto garantiza la escalabilidad del proyecto hacia un asistente virtual completo, alineado con las estrategias de comunicación y servicio digital de la compañía.

2. Preparación de Datos

Extraccion de Informacion General

El proceso de adquisición de información se realizó mediante scripts de web scraping que automatizan la recolección y análisis del contenido textual desde una lista curada de URLs del sitio oficial de Colgate-Palmolive.

Para extraer el contenido institucional —como las secciones de "Quiénes somos", "Historia", "Sostenibilidad", "Informes de gestión", "Noticias" y "Contacto"— se emplearon exclusivamente las librerías requests y BeautifulSoup, enfocadas en la descarga estática de HTML. Estas herramientas permitieron limpiar el contenido visible, descartando elementos irrelevantes como menús, scripts o publicidad.

En paralelo, se incorporó compatibilidad con archivos PDF mediante la librería pdfplumber, utilizada para procesar documentos como informes o reportes descargados desde el sitio. El texto fue extraído página por página y consolidado para su posterior análisis.

Todo el scraping se llevó a cabo bajo principios éticos, sin realizar ataques automatizados masivos ni violar restricciones de acceso. Se priorizó la recolección textual con valor semántico, con miras a alimentar una base de conocimiento útil para el sistema de preguntas y respuestas.

Extracción estructurada de productos

Además del contenido institucional, se implementó un segundo flujo especializado para el catálogo de productos Colgate y Palmolive, cuyo objetivo fue obtener información detallada de cada artículo en un formato estructurado (tipo JSON).

Este proceso sí requirió la automatización de la navegación por las distintas páginas del catálogo utilizando Selenium, ya que muchas de estas secciones dependen de carga dinámica. Se recopilaron atributos clave como nombre del producto, beneficios, ingredientes y categorías.

Cada producto fue almacenado como un objeto JSON con campos estandarizados, lo cual facilitará en etapas posteriores la generación de respuestas más precisas y la incorporación de funcionalidades avanzadas, como búsqueda por atributos o comparación entre productos.

Ejemplo de Objeto JSON Extraído


```
{
  "nombre": "Seda Dental Colgate Total Encerado 25 metros",
  "descripcion": "Complementa tu rutina de higiene personal con el Hilo Dental Colgate
Total Encerado.",
  "imagen": "https://vpxmshare.colgatepalmolive.com/JPEG_1500/vBu3Rx2Zx_8xj6V.jpg",
  "url_detalle":
"https://www.colgate.com/es-co/products/specialty/hilos-dentales-total-encerado",
  "sku": 7891024183182,
  "categoria": "Hilos dentales",
  "marca": "Colgate Total",
  "beneficios": ["Prevenir caries", "Prevenir sarro", "Combatir placa"],
  "descripcion_larga": "Mejora tu rutina diaria de higiene bucal con el Hilo Dental
Colgate Total Encerado, diseñado para eliminar eficazmente la placa entre los
dientes.",
  "faqs": "[{\"pregunta\": \"Ingredientes activos\", \"respuesta\": \"No aplica\"}]",
  "tiendas": "[{\"tienda\": \"FarmatodoCO\", \"disponibilidad\": \"Agotado\",
\"precio\": null}]"
},

{
  "nombre": "Seda Dental Colgate Carbon 25 metros",
  "descripcion": "El hilo dental Colgate Natural Extracts Carbón ayuda a remover
impurezas entre los dientes.",
  "imagen": "https://vpxmshare.colgatepalmolive.com/JPEG_1500/v1lGGoF44kEKDF.jpg",
  "url_detalle":
"https://www.colgate.com/es-co/products/specialty/hilos-dentales-carbon",
  "sku": 7509546671536,
  "categoria": "Hilos dentales",
  "marca": "Colgate Natural Extracts",
  "beneficios": ["Prevenir caries", "Prevenir sarro", "Combatir placa"],
  "descripcion_larga": "Mejora tu higiene bucal con el Hilo Dental Colgate Natural
```

```
Extracts Carbón, enriquecido con ingredientes naturales.",
  "faqs": "[{\"pregunta\": \"¿Cómo usar el hilo dental?\", \"respuesta\": \"Usar el
hilo dental después del cepillado para mejores resultados.\"}]",
  "tiendas": "[{\"tienda\": \"ExitoCO\", \"disponibilidad\": \"Disponible\",
\"precio\": null}]"
}
```


Este enfoque permitió capturar información semánticamente rica, con descripciones, beneficios y categorías asociadas, que posteriormente se integraron en la base de conocimiento del sistema Q&A.

La estructura de estos registros facilitará en etapas posteriores la generación de respuestas más personalizadas o filtros según intereses específicos del usuario final.

Extracción de contenido desde YouTube

También se integró un tercer canal de recolección de información: el perfil oficial de YouTube de Colgate Colombia (youtube.com/colgatecolombia). Para ello se utilizó la API de YouTube Data v3, que permitió extraer metadatos públicos asociados a los videos publicados en el canal.

El proceso de scraping se realizó a través de un script automatizado que empleó la clave de API para consultar el canal, obtener las IDs de los videos más recientes y, posteriormente, recuperar información detallada de cada uno. Los atributos extraídos incluyeron:

- Título del vídeo
- Descripción
- Fecha de publicación
- Enlace directo
- Tags (cuándo disponibles)

Esta información también fue consolidada en un formato estructurado (JSON) y añadida a la base de conocimiento del sistema de preguntas y respuestas, con el objetivo de enriquecer las respuestas del usuario con referencias audiovisuales pertinentes.

Limpieza, normalización y consolidación del conocimiento

Una vez recopilados los datos, se realizó un proceso de limpieza y normalización orientado a garantizar la coherencia semántica y la compatibilidad con los modelos de lenguaje.

Limpieza del texto:

- Eliminación de etiquetas HTML, caracteres especiales, saltos de línea y espacios redundantes.
- Corrección de codificación (UTF-8) y normalización de acentos.
- Exclusión de fragmentos repetidos o sin valor informativo (menús, pie de página, botones, etc.).

Unificación del formato:

Todo el contenido textual (informativo + productos) se almacenó en archivos .json y .txt, con metadatos que indican su origen (url, fecha_scraping, categoría).

Se consolidó una base de conocimiento unificada, que combina texto de la empresa con catálogos de productos, asegurando una representación equilibrada de la empresa y su portafolio.

Segmentación semántica (chunking):

El resultado del scraping se planea segmentar en fragmentos de entre 150 y 200 palabras, con una superposición de 20-30 palabras para mantener la continuidad de contexto.

Cada fragmento se etiquetará con un identificador único y su fuente original, facilitando la trazabilidad durante las consultas del asistente.

El resultado será un corpus limpio y estructurado, compuesto por fragmentos textuales y registros de productos, listo para ser utilizado como contexto base el sistema de preguntas y respuestas.

3. Modelado: Modelo de embedding, LLM, base de datos vectorial y el diseño del prompt.

Modelo de Embedding

Se contemplan modelos de embedding ligeros y ampliamente soportados por frameworks como LangChain, tales como all-MiniLM o BGE, que convierten texto en vectores numéricos de alta densidad. La elección final estará alineada con el modelo de lenguaje que se seleccione para el Módulo 2 (por ejemplo, Gemini 2.5 Flash o modelos ligeros ejecutados localmente mediante Ollama), con el objetivo de garantizar compatibilidad y coherencia semántica en el proceso de recuperación aumentada por búsqueda (RAG).

Modelo de LLM

Aún no se ha definido un LLM definitivo, ya que el sistema se encuentra en fase exploratoria. Actualmente se están realizando pruebas comparativas con distintos modelos, tanto locales como accesibles vía API, priorizando opciones de código abierto o de bajo costo que permitan despliegue controlado. Entre los modelos en evaluación se encuentran:

- **Gemini 2.5 Flash** (Google AI Studio): modelo alojado en la nube, rápido y eficaz para tareas de respuesta a preguntas, con ventana de contexto de +1M tokens de entrada y 65k tokens de salida según la documentación de Google. Esto favorece el manejo de múltiples fuentes simultáneas y prompts largos.

- **Gemma 3 (1B)** (vía Ollama): modelo open-source optimizado para entornos locales y bajo consumo de memoria, con ventana de contexto de 32K tokens en el tamaño 1B (los

tamaños 4B/12B/27B soportan 128K). Adecuado para prototipado eficiente en local.

- **Qwen 3 (1.7B)** (también vía Ollama): modelo multilingüe con buen rendimiento en tareas de comprensión y generación, con ventana de contexto de 32K tokens en la variante 1.7B.

Todos los modelos serán evaluados conforme a criterios de precisión, velocidad de inferencia, capacidad de comprensión contextual, tamaño de la ventana de contexto y adaptación al dominio específico de Colgate-Palmolive. En entregas futuras, se seleccionará un modelo final que permita implementar un sistema RAG robusto, posiblemente con mecanismos de afinamiento (prompt tuning o fine-tuning) para mejorar la precisión de las respuestas.

Base de Datos Vectorial

De manera preliminar, se está explorando el uso de FAISS, por su simplicidad, soporte dentro de LangChain y facilidad de implementación en entornos locales. No obstante, también se analizarán alternativas como Pinecone o Weaviate en caso de que el proyecto requiera mayor escalabilidad o funcionalidades avanzadas de búsqueda y filtrado.

La selección final dependerá del volumen de datos a indexar, los requerimientos de desempeño y la compatibilidad con el modelo de *embedding* y el LLM que se adopte en la siguiente entrega. En general, la decisión buscará un equilibrio entre facilidad de uso, costo computacional y capacidad de integración con el resto de la arquitectura.

Prompt engineering

El componente de generación de respuestas del sistema se desarrolló sobre un único prompt maestro, diseñado para interactuar directamente con el contexto que es el resultado de la concatenación de los diferentes archivos obtenidos del proceso de scraping en un archivo de texto plano .txt.

Este enfoque busca simular el comportamiento inicial de un asistente de preguntas y respuestas (Q&A) basado en memoria textual completa, sin aplicar todavía un mecanismo de recuperación semántica por similitud (RAG).

Orquestación con Langchain

La implementación de LangChain en este proyecto cumple la función de orquestador semántico entre los contextos y la interfaz de usuario. En términos generales, LangChain actúa como la capa intermedia que estructura, controla y gestiona el flujo de comunicación entre el modelo seleccionado y las consultas provenientes del usuario. Gracias a su diseño modular, LangChain permite estandarizar la forma en que se construyen los mensajes, se transmiten las instrucciones y se obtienen las respuestas, lo que garantiza consistencia y trazabilidad en el procesamiento del lenguaje natural.

En este proyecto, LangChain se configura mediante la clase ChatOllam, que permite inicializar el modelo con parámetros específicos como la temperatura, la validación de inicialización y el

número máximo de tokens generados. Esta configuración facilita ajustar el comportamiento del modelo de acuerdo con el tipo de interacción deseada, ya sea respuestas más creativas o más precisas.

La interacción se ejecuta a través del método invoke, que se encarga de enviar la instrucción compuesta al modelo, recibir la respuesta procesada y devolverla en un formato estandarizado. Durante este proceso, LangChain abstrae la complejidad de la comunicación con Ollama, transformando la llamada al modelo en una operación controlada y reproducible. Además, su integración permite registrar el tiempo de ejecución de cada consulta, lo cual aporta una métrica de desempeño útil para evaluar la eficiencia del modelo en distintos escenarios de uso.

4. Resultados

Interfaz en Streamlit y pruebas

La implementación de Streamlit tiene como objetivo ofrecer una interfaz visual e interactiva que facilite la comunicación entre el usuario y el modelo de lenguaje. Esta herramienta cumple la función de capa de presentación dentro del sistema, encargada de gestionar las entradas del usuario, mostrar los resultados del modelo y mantener la continuidad del flujo conversacional en pantalla.

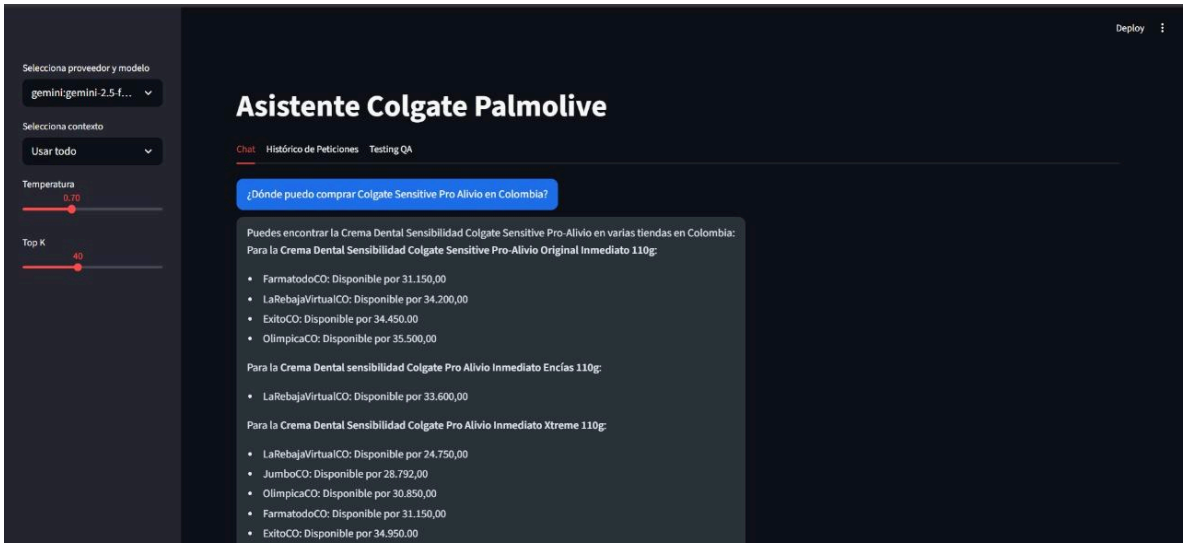

Figura 2. Interfaz de streamlit pestaña Chat.

En la aplicación desarrollada, Streamlit administra tanto las preguntas ingresadas manualmente como un conjunto de veinte consultas predeterminadas que pueden ejecutarse de forma secuencial (Testing QA). La aplicación incorpora una interfaz tipo chat, en la que cada interacción se conserva en pantalla mediante contenedores persistentes. De esta forma, las preguntas y respuestas permanecen visibles en orden cronológico, reproduciendo la dinámica de una conversación continua. Además, cada vez que se envía una consulta, la interfaz muestra un indicador de progreso y registra el tiempo de respuesta, el cual se presenta automáticamente en segundos o minutos según la duración del proceso. Este mecanismo de retroalimentación inmediata mejora la experiencia del usuario y permite evaluar de manera clara el rendimiento del modelo durante la sesión.

Finalmente, la interfaz permite seleccionar el modelo y el contexto de ejecución, además de configurar parámetros como Temperatura y Top-K. También registra el historial de ejecuciones, almacenando información relevante de las consultas como: timestamp, modelo utilizado, valores de temperatura y top_k, mensajes del usuario y del modelo, así como el número de tokens de entrada y salida.

Figura 3. Interfaz de streamlit pestaña de Histórico de peticiones.

En la sección Testing QA se automatizan las pruebas utilizando un conjunto de 30 preguntas base, definidas en un archivo CSV que contiene las columnas pregunta y respuesta esperada. Estas preguntas se integran en una tabla que, además de las columnas mencionadas, incluye la respuesta obtenida por el asistente. A continuación, se presentan los resultados correspondientes a tres preguntas seleccionadas aleatoriamente.

Pregunta	Respuesta Esperada	Respuesta Obtenida	Tiempo

¿Qué marcas de cuidado de la piel premium adquirió Colgate?	PCA Skin, EltaMD y Filorga.	Colgate adquirió las siguientes marcas de cuidado de la piel premium: * **PCA Skin** * **EltaMD** * **Filorga**	4.56 seg
¿Qué producto lanzó Colgate para dientes sensibles a temperaturas extremas?	Colgate Sensitive Pro Alivio Inmediato Xtreme Temperatures.	Colgate lanzó la **Crema Dental Sensibilidad Colgate Pro Alivio Inmediato Xtreme 110g** para dientes sensibles a temperaturas extremas.	3.48 seg
¿Qué relación tuvo Palmolive con Peet Brothers?	Se fusionaron en 1926 formando Palmolive-Peet Company.	Palmolive y Peet Brothers se fusionaron en 1926 para formar la compañía Palmolive-Peet Company.	3.28 seg

Tabla 2. Test QA, muestra 3 preguntas aleatorias.

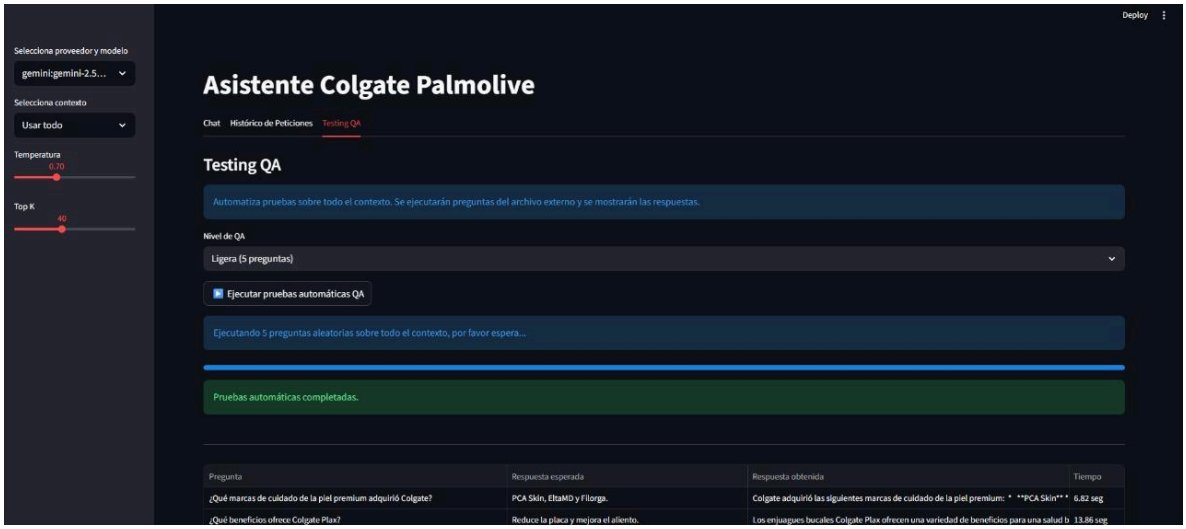

Figura 4. Interfaz de streamlit pestaña de Testing QA.

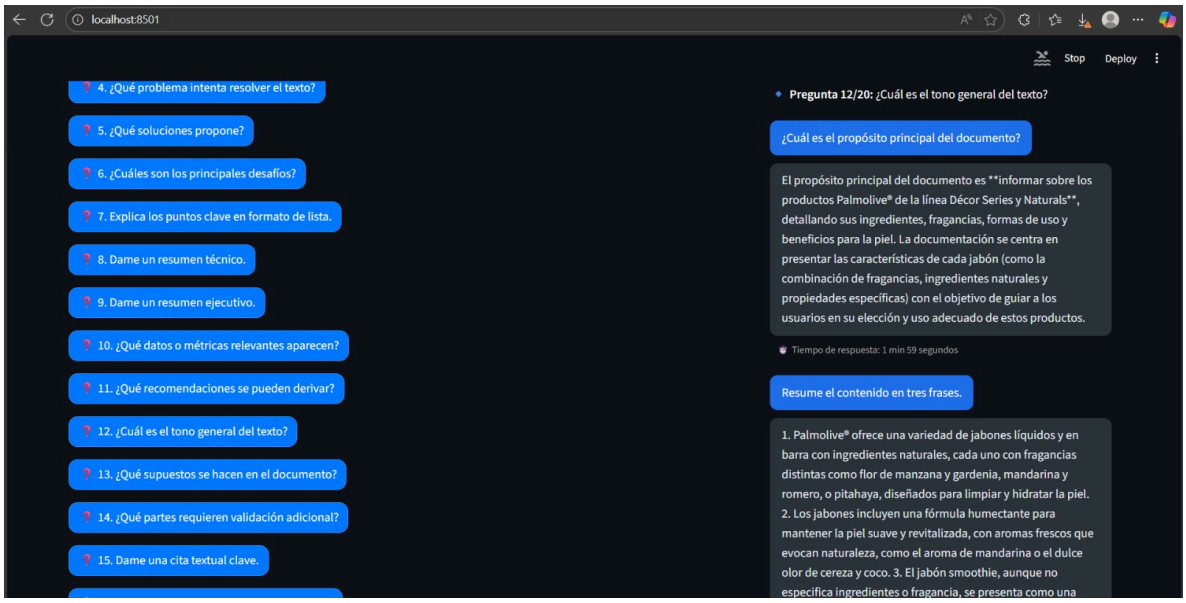

Figura 5. Prueba preliminar con Interfaz previa de streamlit probando Testing QA con modelo qwen3:1.7b.

Link del GitHub: https://github.com/SorenAcevedo/tecnicas_avanzadas_llm

5. Referencias

LangChain. (2024). *Framework for developing applications with LLMs*. https://www.langchain.com/

Streamlit. (2024). *The fastest way to build and share data apps*. https://streamlit.io/

Selenium Documentation. (2025). *Selenium WebDriver for browser automation.* https://www.selenium.dev/

Requests Documentation. (2025). Requests: HTTP for Humans. https://docs.python-requests.org/

BeautifulSoup Documentation. (2025). Beautiful Soup: HTML/XML parsing library. https://www.crummy.com/software/BeautifulSoup/bs4/doc/