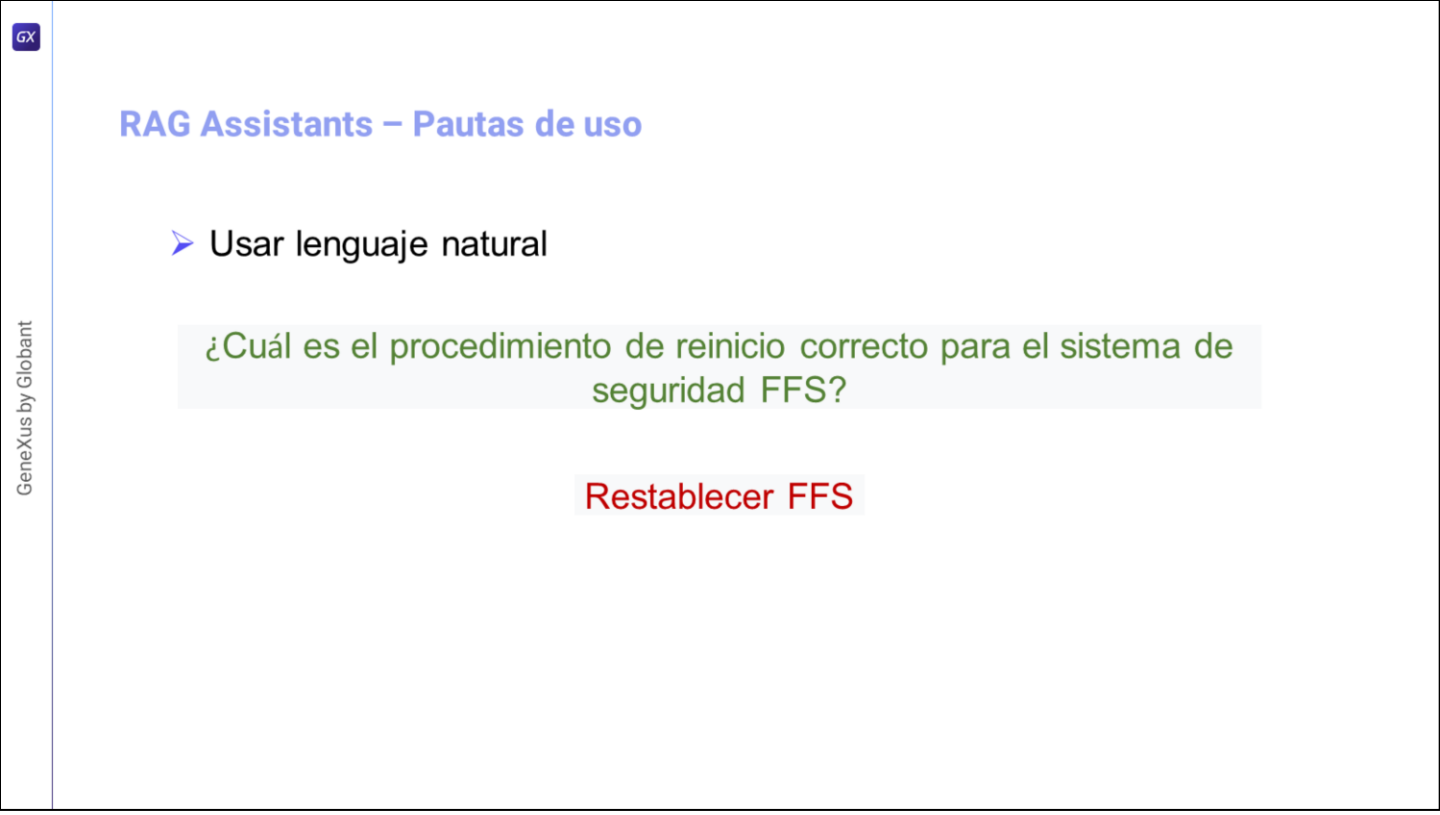
GeneXus by Globant

RAG Assistants – Pautas de uso y Configuración

Alejandra Caggiano

Avanzando un poco más, ahora que ya sabemos cómo crear y testear un asistente, vamos a ver algunas pautas sobre cómo hacer preguntas correctas para obtener información que formó parte de la ingesta de documentos de un determinado Asistente RAG.

Veamos entonces algunas consideraciones generales:

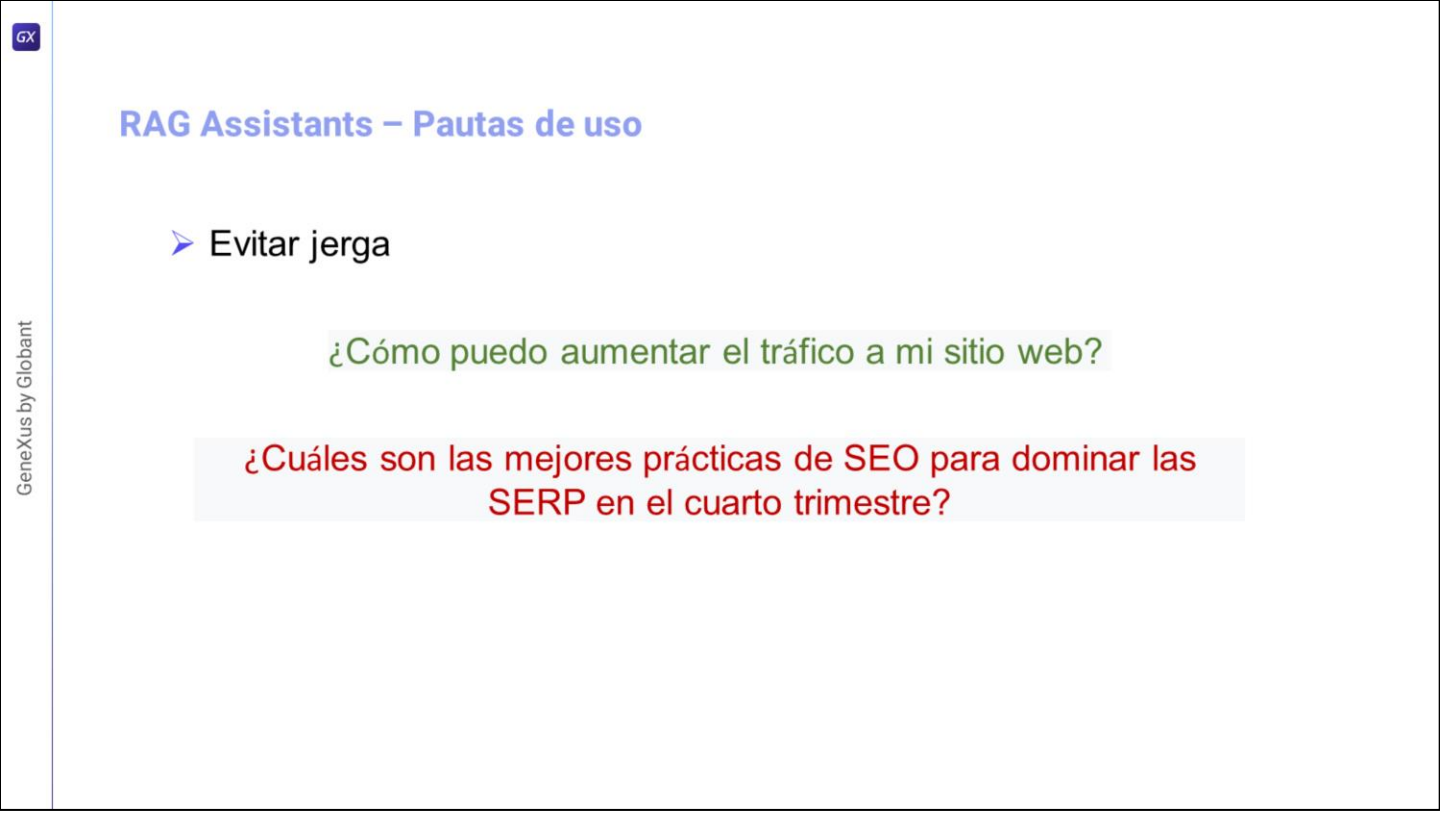


La primera consideración es **utilizar lenguaje natural:** Esto significa escribir preguntas con naturalidad, tal como si se estuviera hablando con una persona.

No importa si se comete algún error ortográfico o se utilizan sinónimos. Lo importante ese transmitir correctamente la intención.

Para poner un ejemplo, una pregunta correcta puede ser: ¿Cuál es el procedimiento de reinicio correcto para el sistema de seguridad FFS?

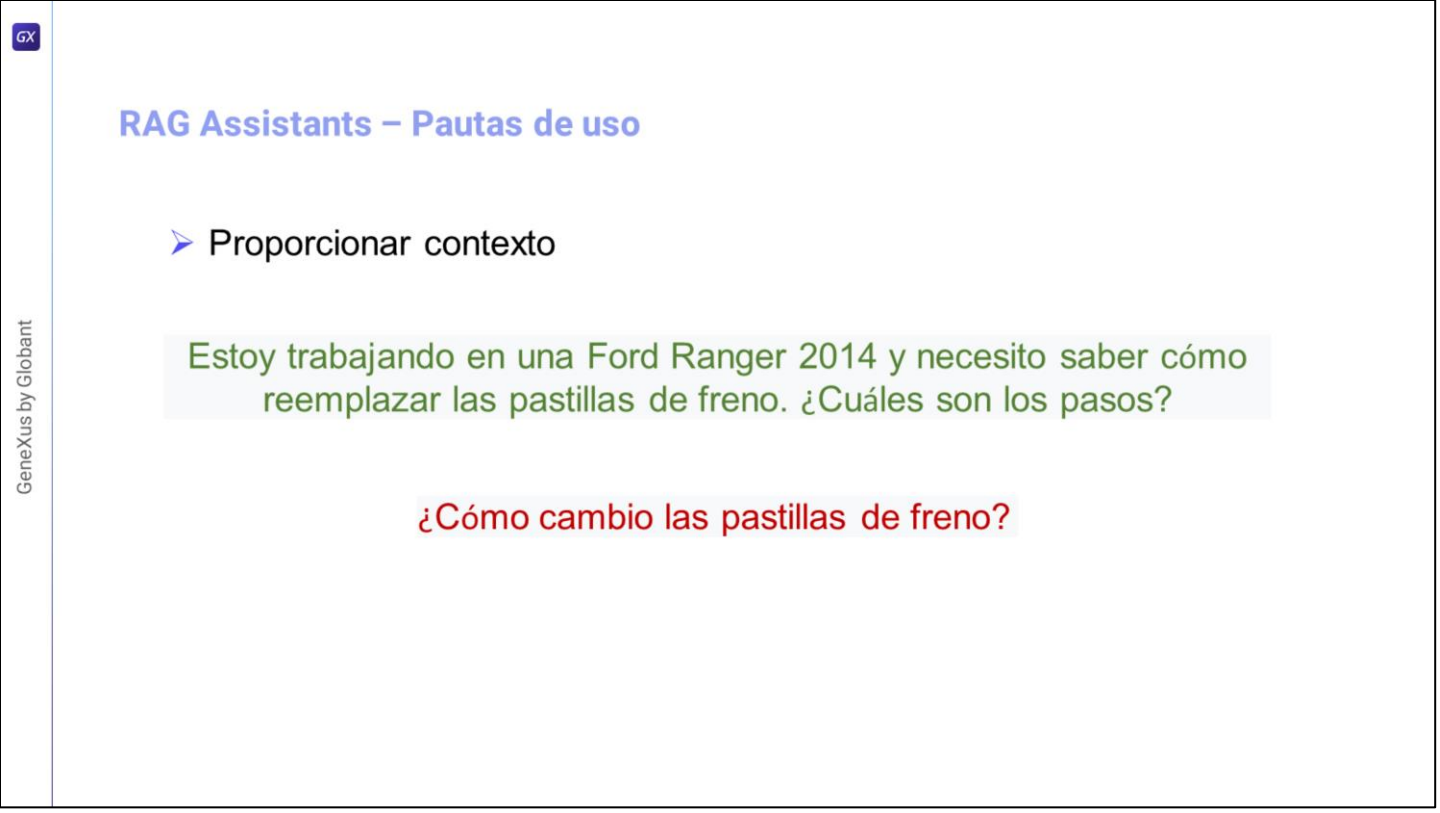
Mientras que una expresión incorrecta sería escribir simplemente: Restablecer FFS



Bien. Otra consideración a tener en cuenta es **evitar la jerga**: Aunque el asistente puede comprender términos técnicos, un lenguaje más simple a menudo puede proporcionar resultados más precisos.

Como ejemplo, una expresión correcta sería: ¿Cómo puedo aumentar el tráfico a mi sitio web?

Mientras que una expresión incorrecta sería: ¿Cuáles son las mejores prácticas de SEO para dominar las SERP en el cuarto trimestre?



Otra consideración importante es la de **proporcionar un contexto**: Agregar un poco de contexto o antecedentes a la consulta puede ayudar al asistente a comprender mejor el alcance de la solicitud.

Una expresión correcta sería: Estoy trabajando en una Ford Ranger 2014 y necesito saber cómo reemplazar las pastillas de freno. ¿Cuáles son los pasos?

Mientras que no sería bueno simplemente consultar: ¿Cómo cambio las pastillas de freno?

Otra pauta recomendada es utilizar un **lenguaje claro y conciso**: Es importante ser específico y articular claramente lo que se está buscando para recibir información más precisa y relevante. Evitar la ambigüedad ayuda al asistente a proporcionar respuestas precisas.

Un ejemplo correcto sería consultar: ¿Cuál es el proceso para que los ciudadanos uruguayos soliciten una visa de turista para Japón?

En cambio, no sería recomendado simplemente consultar: ¿Cómo viajo a Japón?

RAG Assistants – Pautas de uso

- Realizar una sola pregunta a la vez

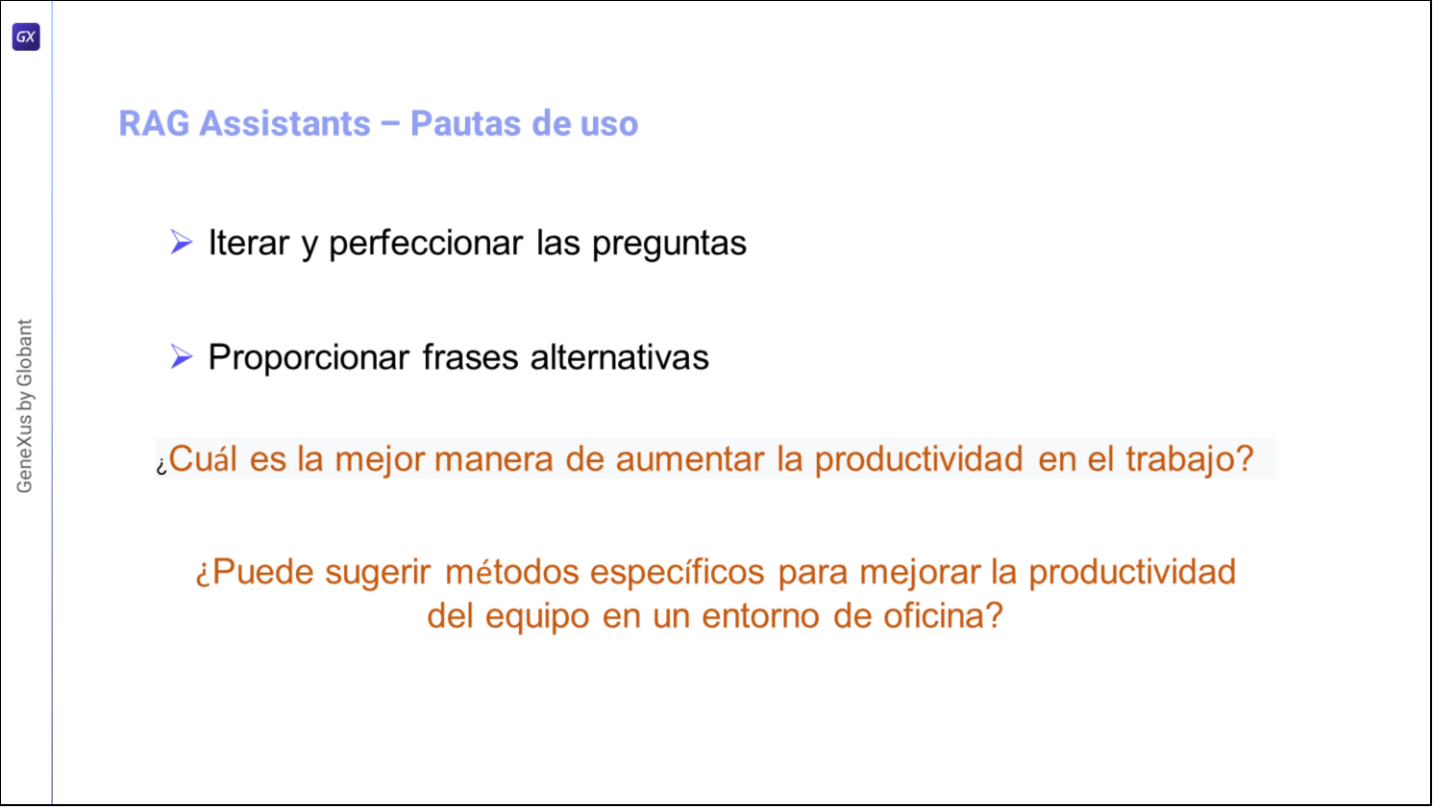
¿Cuál es la temperatura promedio en París en junio? ¿Cuáles son algunas de las atracciones turísticas populares allí?

¿Cómo es el clima en París y qué debo hacer durante mi visita?

Otra buena pauta **es hacer una sola pregunta a la vez:** Si debemos hacer una consulta compleja, es bueno considerar dividirla en partes más simples.

Un ejemplo correcto sería: ¿Cuál es la temperatura promedio en París en junio? ¿Cuáles son algunas de las atracciones turísticas populares allí?

En cambio, no sería lo mejor consultar: ¿Cómo es el clima en París y qué debo hacer durante mi visita?



Otra buena consideración es **iterar y perfeccionar las preguntas**: Si la respuesta inicial es diferente de lo que se necesita, es bueno rearmar la pregunta según la respuesta brindada por el asistente.

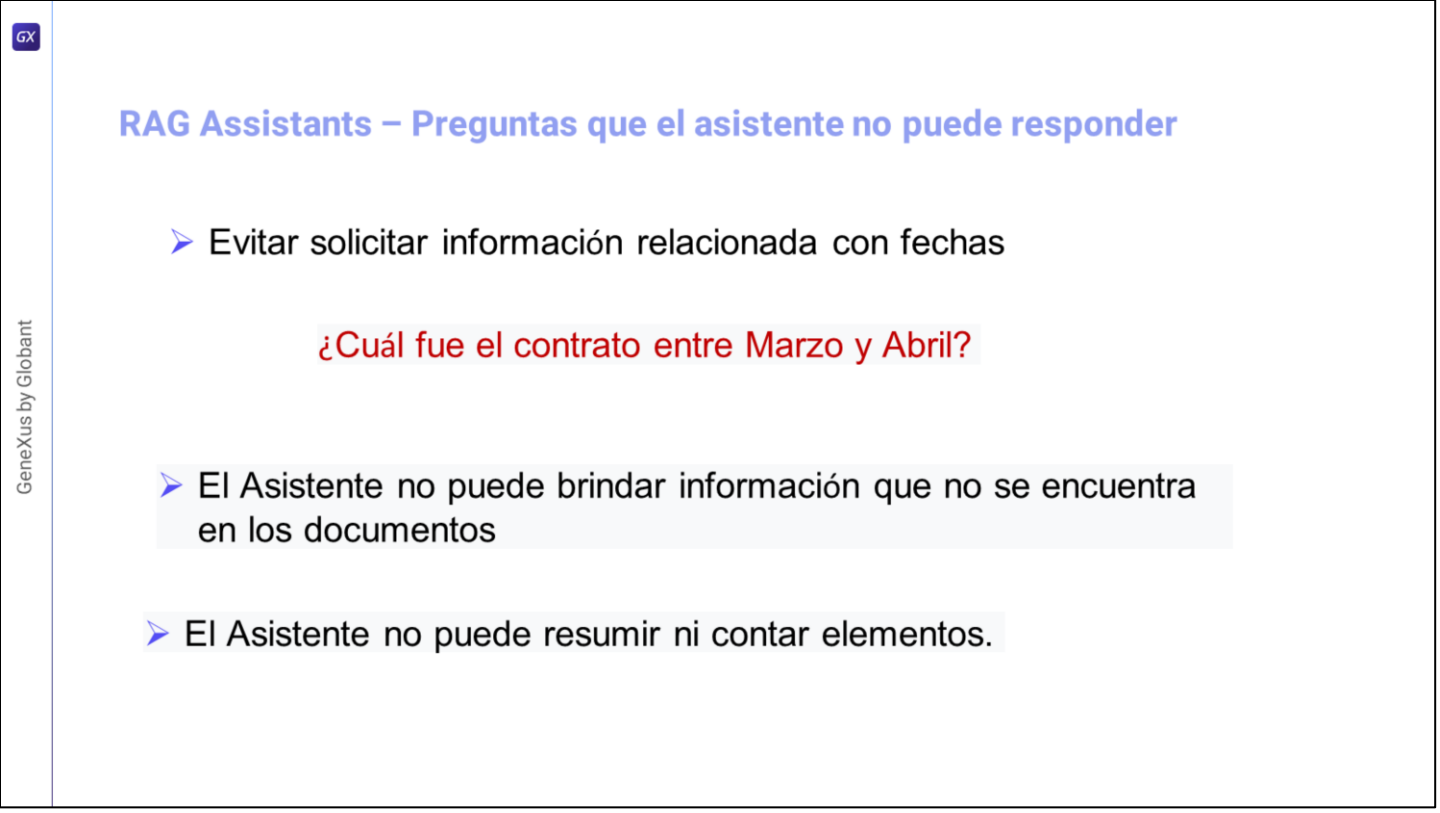
También es bueno **proporcionar frases alternativas**: Esto significa probar con diferentes formas de formular una consulta, incluyendo sinónimos y variaciones con el fin de explorar la compresión del asistente.

Veamos como ejemplo la siguiente consulta:

¿Cuál es la mejor manera de aumentar la productividad en el trabajo? Si el asistente no comprende la pregunta, podría brindar como respuesta, estrategias generales sobre productividad en lugar de estrategias específicas en el lugar de trabajo.

Entonces se puede reformular de la siguiente forma:

¿Puede sugerir métodos específicos para mejorar la productividad del equipo en un entorno de oficina? Esta reformulación, entonces, aclara que la atención se centra en la productividad del equipo en un entorno profesional, y no en consejos de productividad individuales.



Bien. Así como acabamos de ver pautas y consideraciones de buen uso a la hora de comunicarnos con un asistente, es importante tener en cuenta también que **hay ciertas preguntas que el asistente no puede responder.**

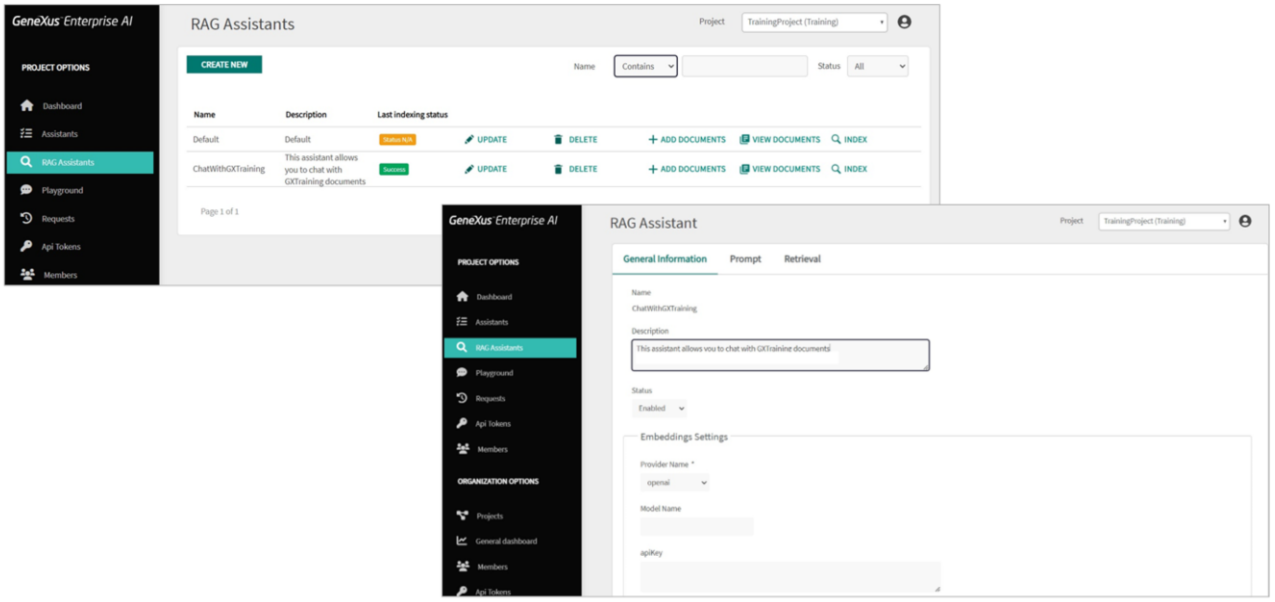
Por ejemplo, es importante evitar pedir información relacionada con fechas. Algo de este estilo no sería recomendable: “¿Cuál fue el contrato entre marzo y abril?”.

No se espera que las comparaciones o el contraste de información den buenos resultados;

Otra consulta que el asistente no puede responder es brindar información que no forma parte de los documentos que se utilizaron para darle contexto al Asistente RAG.

No puede resumir ni contar elementos, y tampoco puede procesar textos que aparezcan en imágenes.

RAG Assistants – Configuración



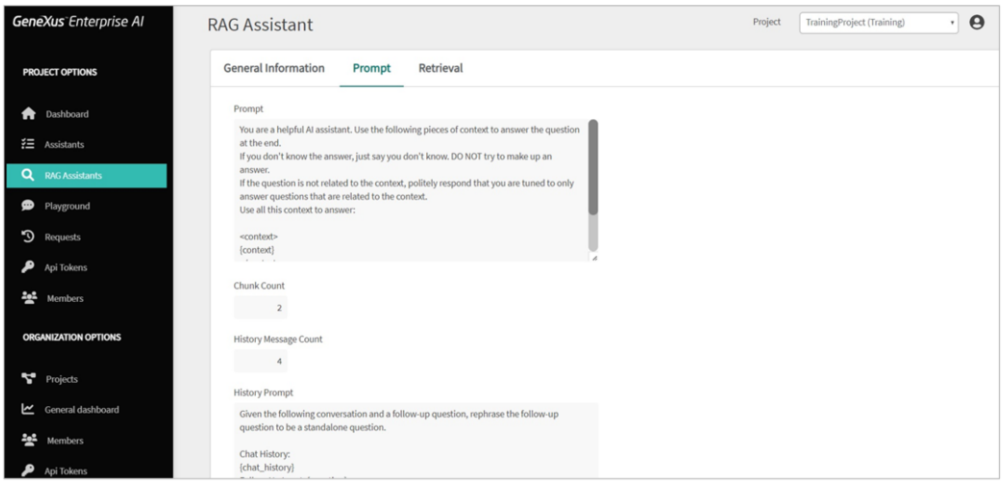
Bien. Ahora que ya sabemos las consideraciones a tener en cuenta para tener una buena comunicación con un asistente RAG, vamos a profundizar un poco más en su configuración.

Como ya hemos visto anteriormente, toda la interacción con el componente de Búsqueda y Chat se configura a través de la sección RAG Assistants. Se crea un asistente RAG predeterminado durante la inicialización y luego se puede modificar o crear otros nuevos para cambiar su comportamiento.

Una vez creado, accedemos a sus opciones mediante el botón Update, y ya sabemos que el conjunto de estas opciones está organizado en tres secciones:

- La sección **General Information**, donde se indicar detalles que permiten comprender rápidamente las capacidades y características del asistente.
- La sección **Prompt**: donde se indica información sobre cómo está configurado el asistente para recibir y procesar consultas de los usuarios.
- Y la sección **Retrieval**: donde se Bespecifica cómo se recupera la información.

RAG Assistants – Configuración



Vayamos a la sección Prompt

Como ya sabemos, aquí se pueden configurar instrucciones que orienten al asistente sobre cómo abordar y responder preguntas. Estas instrucciones establecen pautas claras para que el asistente brinde respuestas relevantes y útiles basadas en el contexto proporcionado. El valor predeterminado es este que estamos viendo.

Se deben mantener configuradas las variables de contexto y pregunta porque serán reemplazadas con la información asociada antes de la interacción.

Bien. Esta siguiente opción define cuántos fragmentos se recuperan para aumentar el contexto.

La opción Cantidad de mensajes del historial establece la cantidad de mensajes históricos que se tienen en cuenta en la conversación.

Esto es útil para rastrear el historial de interacción y comprender el contexto recopilado en la conversación. Hay que considerar que este valor se refiere a la pregunta del usuario final y la respuesta asociada. Es decir, si está establecido en 4, significa que se está interesado en considerar las últimas 4 interacciones:

•vPregunta anterior

• Respuesta anterior

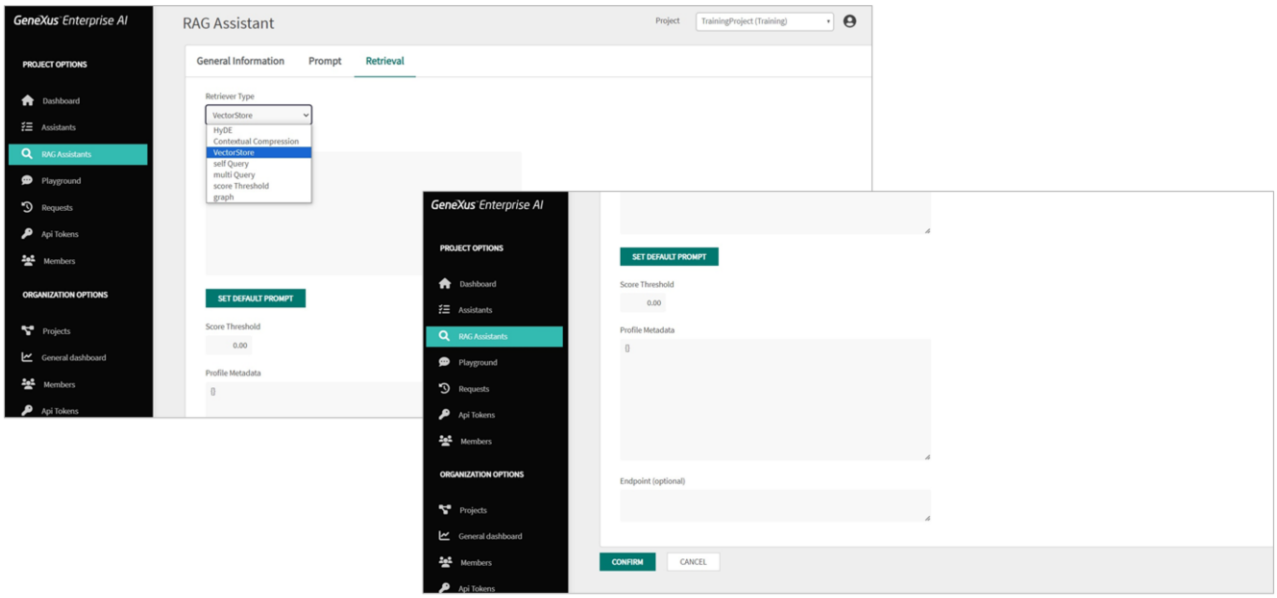
• Última pregunta

• y Última respuesta

El valor mínimo que puede tomar es 0, lo que indica que el historial de conversaciones no es de interés.

En cuanto a las opciones de configuración del LLM, abarcan la configuración del modelo utilizado por el asistente para generar la respuesta, incluido el proveedor de servicios, el nombre del modelo, la temperatura, el límite máximo de token y otros parámetros que afectan la forma en que se generan las respuestas.

RAG Assistants – Configuración



Pasemos a la solapa de **Recuperación**

Esta sección permite especificar cómo obtener la información enviada al contexto.

Acá se puede indicar el tipo de recuperador utilizado para obtener información. El valor predeterminado es VectorStore,, que utiliza directamente el VectorStore definido sin procesamiento previo adicional.

Los otros valores que puede tomar son los siguientes:

Método de **incorporación de documentos hipotéticos**: Este método utiliza técnicas de incorporación para responder consultas, generar respuestas hipotéticas, incorporarlas en el documento generado y utilizarlo luego como ejemplo final.

Otro posible método de recuperación es la **Compresión Contextual**. Este método Intenta mejorar las respuestas devueltas por las búsquedas de similitud de documentos, teniendo en cuenta de mejor manera el contexto de la consulta.

Otra opción es la **Autoconsulta**. En este tipo de recuperación primero se consulta a sí mismo para recuperar información de filtro basada en la consulta en lenguaje natural. Luego ejecuta una segunda consulta al LLM con la consulta y los filtros aplicados en base a la primera.

También se puede elegir la **Consulta múltiple**. Este tipo de recuperador automatiza el proceso de ajuste rápido mediante el uso de un LLM para generar múltiples consultas desde diferentes perspectivas para una cierta consulta inicial de usuario. Para cada consulta entonces, recupera un conjunto de documentos relevantes y toma la unión única de todas las consultas para obtener entonces un conjunto más grande de documentos potencialmente relevantes.

Luego, la opción **Umbral de puntuación**, utiliza lo que se llama búsqueda recursiva de similitud. Se devolverán todas las coincidencias de preguntas similares según el umbral de puntuación mínimo indicado.

Por último, el tipo de recuperador **Grafico** permite el uso de un enfoque de representación de información basado en gráficos para su recuperación.

Y a continuación, en este cuadro se puede especificar la consulta que se envía al recuperador para buscar información. Esta consulta puede ser una pregunta o una solicitud específica.

Este cuadro **Umbral de puntuación** define el valor mínimo válido para considerar la información como válida cuando es recuperada. Si no hay documentos válidos, no se produce ninguna interacción con el LLM. El valor predeterminado es 0,0.

El cuadro **Metadata del perfil.** Puede contener metadata adicional relacionada con el perfil del recuperador. El valor predeterminado es un objeto vacío, pero en determinados casos conviene poder afinar cada acceso al LLM para obtener el resultado deseado.

Finalmente, en el cuadro **Endpoint** se puede indicar la dirección URL que apunta al servidor o servicio específico donde se alojan los métodos o modelos de recuperación.

GeneXus by Globant

GeneXus™
by Globant

training.genexus.com