

Chatear con documentos – RAG Assistant



Alejandra Caggiano

Hemos visto que GeneXus Enterprise AI permite crear distintos tipos de asistentes de inteligencia artificial. En particular, ya hemos logrado crear una conversación interactiva.

Queremos ahora poder definir un asistente que nos permita chatear con documentos. Para eso, vamos a trabajar con asistentes RAG.

Retrieval Augmented Generation (RAG)

La Recuperación de Generación Aumentada (RAG) es un enfoque que combina la recuperación de información a partir de datos no estructurados y la generación de texto para mejorar el rendimiento en tareas como ser la respuesta a preguntas.

Retrieval Augmented Generation (RAG)

- Data Ingestion
- Retrieval
- Generation
- End user interaction

Este proceso se compone de las siguientes cuatro fases:

- La primera fase es la **ingesta de datos**: Implica cargar varios tipos de documentos, en diferentes formatos y desde múltiples fuentes.
- Luego sigue la fase de **Recuperación**: En esta etapa se inicia el proceso de recuperación de datos, aprovechando la información previamente cargada y organizada. Se realiza una búsqueda selectiva sobre un conjunto de documentos, identificando la información relacionada y reduciendo eficientemente el espacio de búsqueda. Este enfoque garantiza que la atención se centre en la información más relevante y significativa.
- La siguiente fase es la de **Generación**: El foco aquí está en generar respuestas relevantes y contextualmente consistentes. En este proceso, el sistema utiliza la configuración del asistente RAG para saber a qué modelo acceder y con qué parámetros. Este asistente incorpora los elementos necesarios para definir la estrategia de búsqueda, y lograr coherencia y relevancia en el contenido generado.
- La fase final es **la Interacción con el usuario final**: GeneXus Enterprise AI facilita una comunicación fluida y eficiente entre los usuarios finales y los asistentes RAG, completando el ciclo y brindando respuestas a las consultas de manera eficiente.

Retrieval Augmented Generation (RAG)

Ejemplo:

Chatear con documentos de GeneXus Training

Bien. Como planteamos al comienzo, nuestro objetivo ahora es crear un asistente que nos permita chatear con un conjunto de documentos, y lo haremos con documentos de GeneXus Training.

RAG Assistants

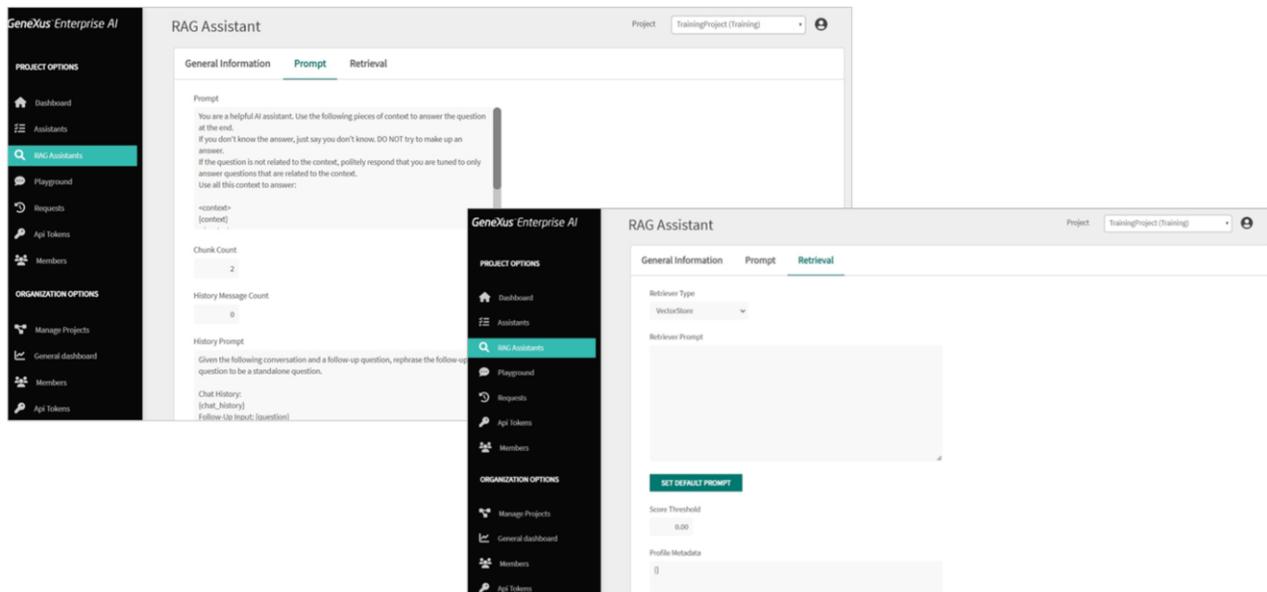
The screenshot displays the GeneXus Enterprise AI interface for managing RAG Assistants. The main view shows a list of assistants with columns for Name, Description, and Last indexing status. A 'CREATE NEW' button is visible. An inset window shows the 'RAG Assistant' configuration form with tabs for General Information, Prompt, and Retrieval. The General Information tab is active, showing fields for Name (ChatWithTraining), Description (This assistant chat with GeneXus training documents), Status (Enabled), and Embeddings Settings (Provider Name: openai, Model Name, apiKey).

Así que ingresamos a la plataforma, seleccionamos el proyecto sobre el cual vamos a trabajar, y en el menú elegimos RAG Assistant. Por defecto, al abrir esta sección vemos un RAG Assistant de nombre Default, que puede ser personalizado, o bien podemos crear nuevos.

Presionamos Create new. Ponemos como nombre ChatWithGXTraining e ingresamos una breve descripción. Presionamos Confirm.

Desde la opción Update podemos personalizar la definición del asistente, según sea necesario. El conjunto de settings se organiza en estas solapas:

RAG Assistants



La Información general, el Prompt, que contiene 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.

Esta opción indica la cantidad de fragmentos que se recuperan para aumentar el contexto.

Luego esta opción de Historial de mensajes 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.

Si este valor está establecido en 4, significa que interesa considerar las últimas 4 interacciones:

El valor mínimo que puede tomar es 0, lo que indica que el historial de conversaciones no es de interés. Cuando el valor es mayor que 0, se utiliza junto con el mensaje indicado en la siguiente opción History Prompt.

RAG Assistants

The screenshot displays the GeneXus Enterprise AI interface for managing RAG Assistants. The top panel shows a list of assistants with columns for Name, Description, and Last indexing status. The 'ChatWithTraining' assistant is highlighted, and its '+ ADD DOCUMENTS' button is circled in blue. The bottom panel shows the 'Search and chat: ChatWithTraining profile' page, which includes a 'Document indexing' section with steps for adding files and metadata, and a table for adding metadata rows.

Name	Description	Last indexing status
Default	Default	Success
ChatWithTraining	This assistant chats with GeneXus trainings documents.	Success

Page 1 of 1

Search and chat: ChatWithTraining profile

Document indexing

Welcome to Search & Chat
GeneXus Enterprise AI will guide you in order to upload a text document so to be able to do Q&A regarding that content

Step 1: Add files for indexing
Select one or more files to upload.

+ Add Files

Step 2: Add document metadata
In order to improve search, add optional document metadata information.

Tags

Name	Value
Add metadata rows with "New row" option	

[New row]

After adding files for indexing, you will be able to try the Search & Chat module.

Luego vemos las opciones para establecer la configuración del modelo utilizado por el asistente para generar la respuesta. Esto incluye el proveedor de servicios, el nombre del modelo, la temperatura, el límite máximo de tokens y otros parámetros que afectan la forma en que se generan las respuestas.

Finalmente, la solapa de Recuperación, que especifica cómo se recupera la información.

Dejamos los valores por defecto.

Bien. Vamos ahora a cargar los documentos. Presionamos Add Documents. El botón Add Files permite realizar la carga de archivos de diversos formatos:

En nuestro ejemplo vamos a cargar un pequeño conjunto de pdfs y docs que corresponden a material de GeneXus training.

Presionamos Add Files

RAG Assistants

The screenshot displays the GeneXus Enterprise AI interface for RAG Assistants. It is divided into two main sections: the upload configuration step and the assistant management view.

Step 1: Add files for indexing
Select one or more files to upload.

File Name	Size	Action
CourseIntroduction_en.pdf	582.27 KB	Cancel
WhatsAppEnterprise_en.docx	232.79 KB	Cancel
GXEnterprise_BackOffice_en.pdf	686.67 KB	Cancel

Step 2: Add document metadata
In order to improve search, add optional document metadata information.

Tags

Name	Value
Add metadata rows with "New row" option	

After adding files for indexing, you will be able to try the Search & Chat module.

RAG Assistants
Project: TrainingProject (Training)

CREATE NEW Name: Contains Status: All

Name	Description	Last indexing status	Actions
Default	Default	Completed	UPDATE DELETE + ADD DOCUMENTS VIEW DOCUMENTS
ChatWithTraining	This assistant chats with GeneXus training documents.	In progress (Indexing file 2 of 2)	UPDATE DELETE + ADD DOCUMENTS VIEW DOCUMENTS

Page 1 of 1

Una vez cargados los documentos presionamos Start upload

Para ver los documentos cargados, cada uno con su detalle, presionamos View Documents

RAG Assistants

The screenshot displays the GeneXus Enterprise AI interface. The top section shows a list of indexed documents for the 'ChatWithTraining' profile. The table below summarizes the data shown:

Name	Extension	Timestamp	Index Status	Action
GXEnterprise_Backoffice_en	pdf	02/05/24 04:38:48,573 PM	Success	DELETE
WhatsGXEnterpriseAI_en	docx	02/05/24 04:38:48,266 PM	Success	DELETE
ConsultasIndustria.es	pdf	02/05/24 04:38:47,920 PM	Success	DELETE

The bottom section shows the metadata for the selected document 'WhatsGXEnterpriseAI_en'. The metadata is organized into two columns:

General	Metadata
ID	Name
1000004f-6929-4943-a044-c1018f8c94d1	WhatsGXEnterpriseAI_en
Organization Name	Project ID
Training	81046361-53c3-44ff-af10-5a76163a36e6
Project Name	Profile Name
TrainingProject	ChatWithTraining
Description	Extension
	docx
Key Name	File Name
upload/1000004f-6929-4943-a044-c1018f8c94d1/document.docx	WhatsGXEnterpriseAI_en.docx
Document URI	Chunks
https://api.gx.enterprise.ai/api/document/1000004f-6929-4943-a044-c1018f8c94d1/document.docx?lang=es&securityToken=eyJ0eXAiOiJKV1QiLCJhbGciOiJIUzI1NiJ9.eyJ1bmludGVudCI6ImFkbG9jL3VudC91bmludGVudC5kb2N1IiwiaWF0IjoiMjAyNC0wNS0yNC00ODM4NDg1NzNPMiwiZXN0IjoiZm9udCJ9	https://api.gx.enterprise.ai/api/chunks/1000004f-6929-4943-a044-c1018f8c94d1/document.docx?lang=es&securityToken=eyJ0eXAiOiJKV1QiLCJhbGciOiJIUzI1NiJ9.eyJ1bmludGVudCI6ImFkbG9jL3VudC91bmludGVudC5kb2N1IiwiaWF0IjoiMjAyNC0wNS0yNC00ODM4NDg1NzNPMiwiZXN0IjoiZm9udCJ9

Seleccionando el nombre del archivo vemos toda la información asociada, pudiendo ver y descargar el archivo desde la URL.

Bien. Ya tenemos creado nuestro RAG Assistant y lo hemos cargado con los archivos correspondientes. Estamos entonteces en condiciones de testarlo.

Lo haremos a continuación desde la opción Playground del menú.

GeneXus[™]
by **Globant**

training.genexus.com