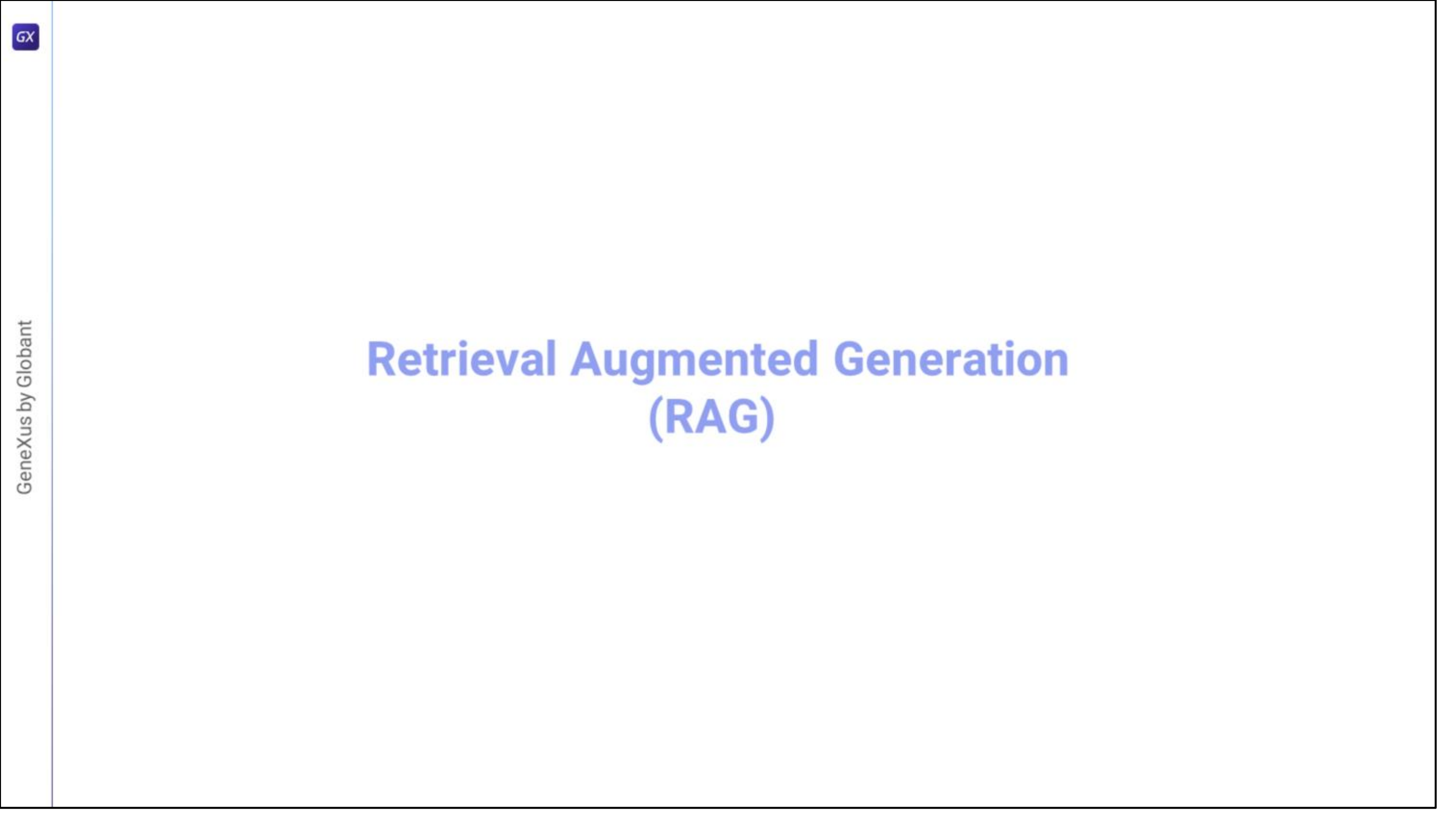
GeneXus by Globant

Chat with documents – RAG Assistant

Alejandra Caggiano

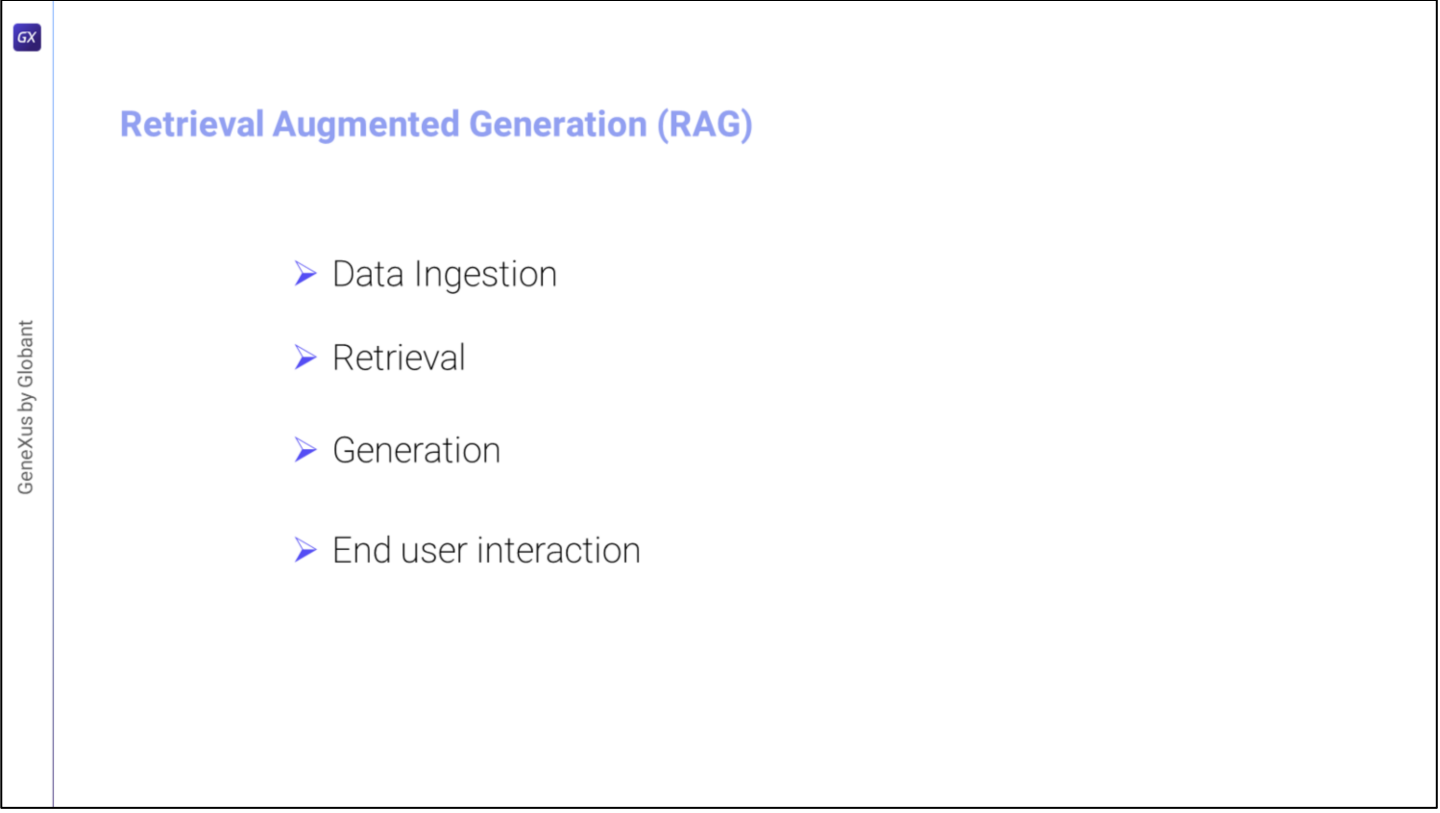
We have seen that GeneXus Enterprise AI allows creating different types of Artificial Intelligence assistants. In particular, we have already succeeded in creating an interactive conversation.

Now we want to be able to define an assistant that allows us to chat with documents. For that, we are going to work with RAG Assistants.



Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) is an approach that combines information retrieval from unstructured data and text generation to improve performance in tasks such as question answering.



GeneXus by Globant

Retrieval Augmented Generation (RAG)

- Data Ingestion
- Retrieval
- Generation
- End user interaction

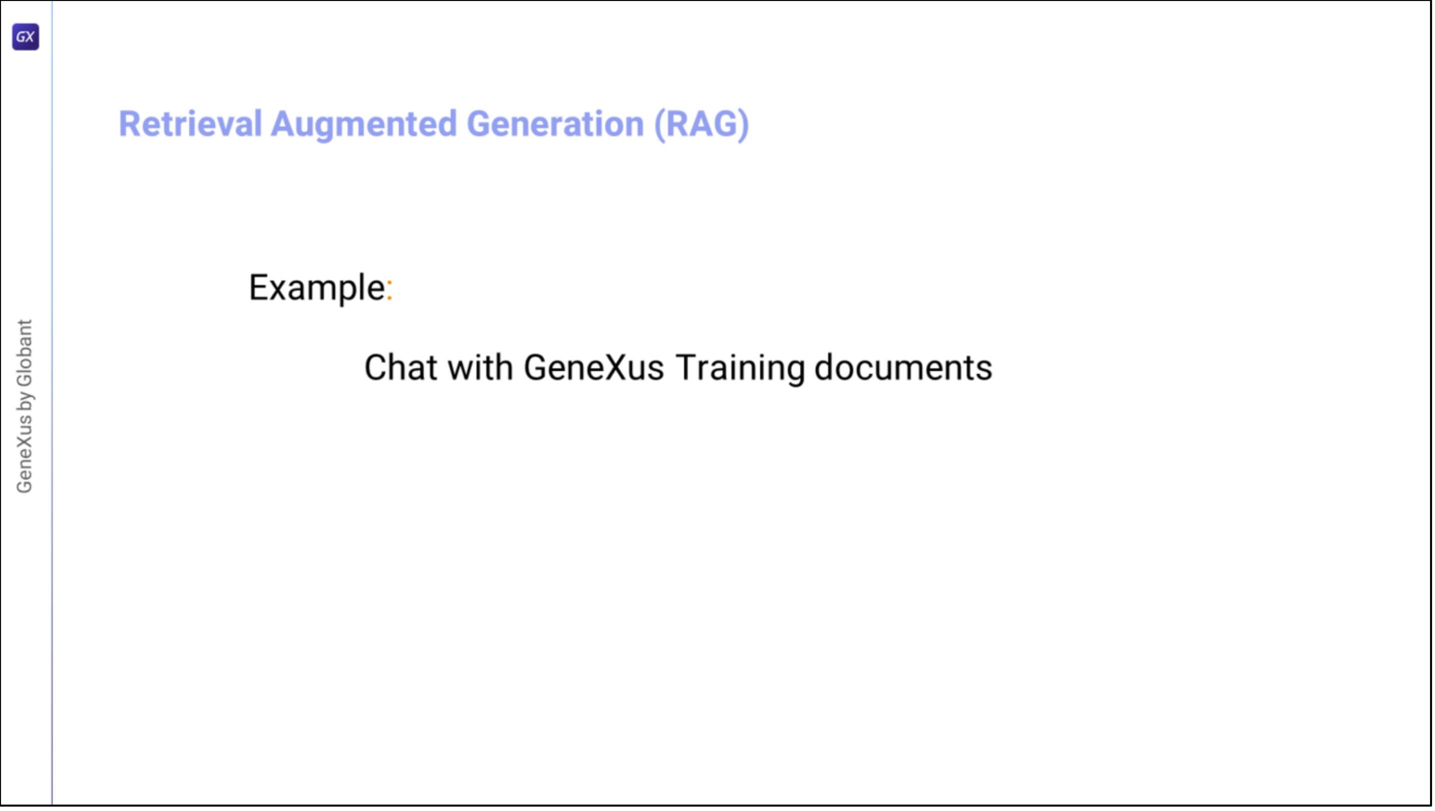
This process consists of the following four phases:

- The first phase is **data ingestion**: It involves loading various types of documents, in different formats and from multiple sources.

- It is followed by the **Retrieval phase:** In this stage, the data retrieval process is started, taking advantage of the previously loaded and organized information. A selective search is performed on a set of documents, identifying related information and efficiently reducing the search space. This approach ensures that focus is placed on the most relevant and meaningful information.

- The next phase is **Generation**: It focuses on generating relevant and contextually consistent answers. In this process, the system uses the RAG Assistant configuration to know which model to access and with which parameters. This assistant incorporates the necessary elements to define the search strategy, and to achieve consistency and relevance in the generated content.

- The last phase is **Interaction with the end user**: GeneXus Enterprise AI facilitates smooth and efficient communication between end users and RAG Assistants, completing the cycle and providing answers to queries in an efficient manner.



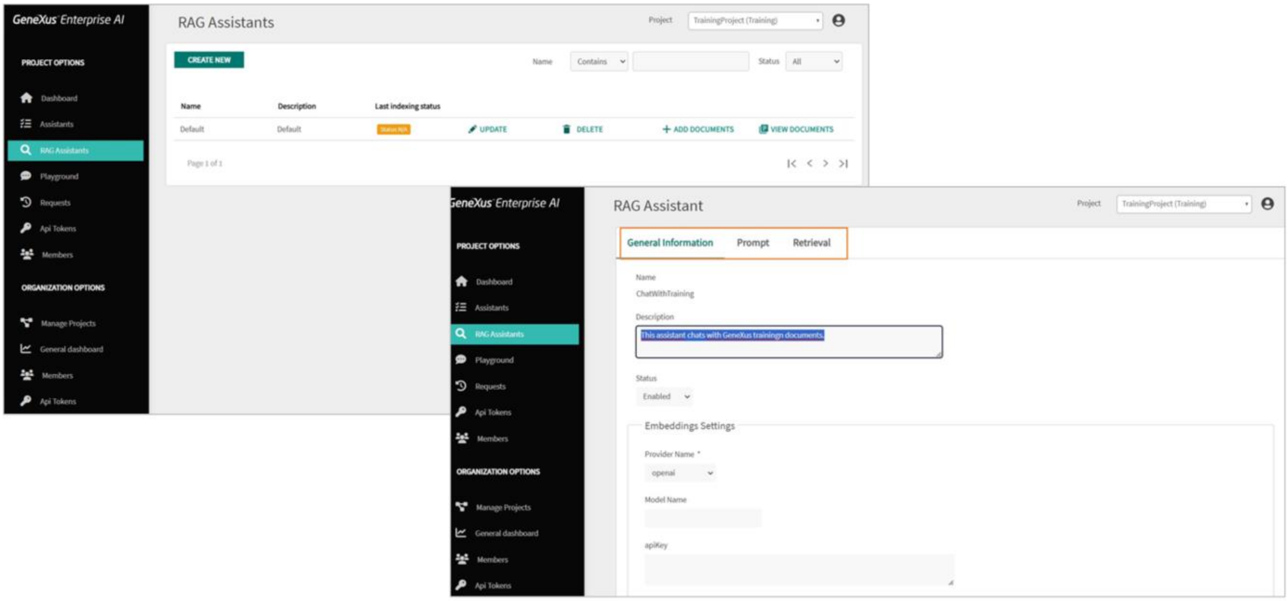
Retrieval Augmented Generation (RAG)

Example:

Chat with GeneXus Training documents

Good. As we said at the beginning, our goal now is to create an assistant that allows us to chat with a set of documents, and we will do it with GeneXus Training documents.

RAG Assistants



Name	Description	Last indexing status				
Default	Default	Status N/A	UPDATE	DELETE	+ ADD DOCUMENTS	VIEW DOCUMENTS

Name
ChatWithTraining

Description
This assistant chats with GeneXus trainingn documents.

Status
Enabled

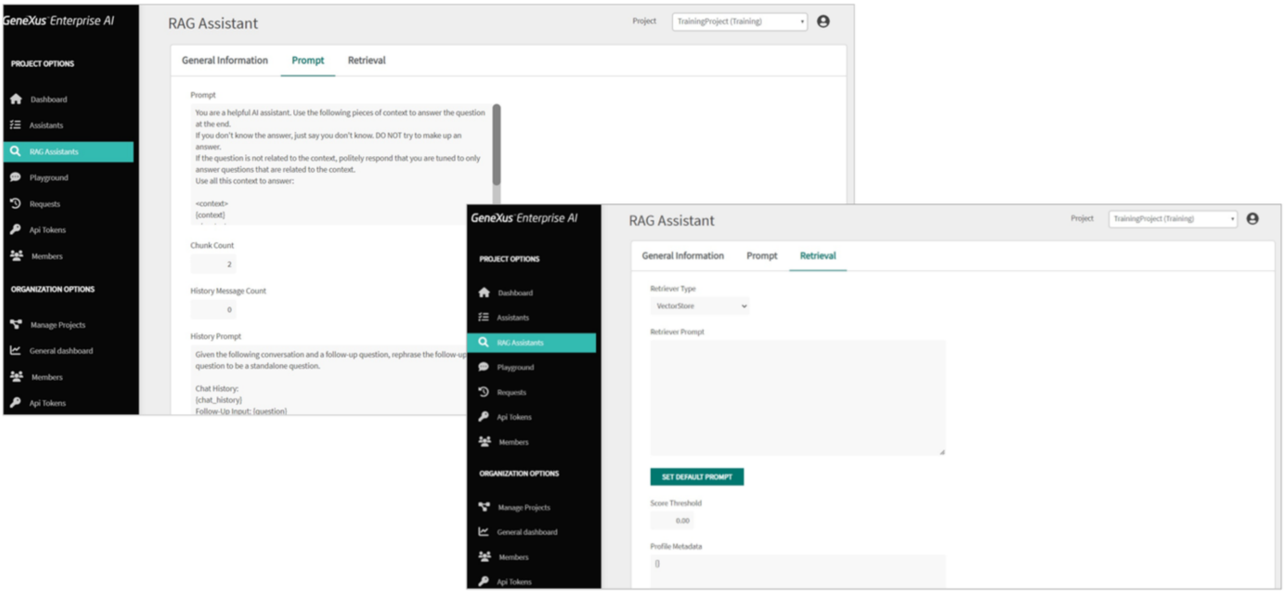
Provider Name *
openai

Model Name

apiKey

So we enter the platform, select the project we are going to work on, and in the menu we choose RAG Assistant. By default, when opening this section we see a RAG Assistant named Default, which can be customized, or we can create new ones. We click on Create new, name it ChatWithGXTraining, and enter a brief description. We Confirm. From the Update option, we can customize the assistant definition as needed. The settings are organized in the following tabs:

RAG Assistants



RAG Assistant — Project: TrainingProject (Training) — Prompt

Prompt

You are a helpful AI assistant. Use the following pieces of context to answer the question at the end.
If you don't know the answer, just say you don't know. DO NOT try to make up an answer.
If the question is not related to the context, politely respond that you are tuned to only answer questions that are related to the context.
Use all this context to answer:

<context>
{context}

Chunk Count: 2

History Message Count: 0

History Prompt

Given the following conversation and a follow-up question, rephrase the follow-up question to be a standalone question.

Chat History:
{chat_history}
Follow-Up Input: {question}

RAG Assistant — Project: TrainingProject (Training) — Retrieval

Retriever Type: VectorStore

Retriever Prompt

Score Threshold: 0.00

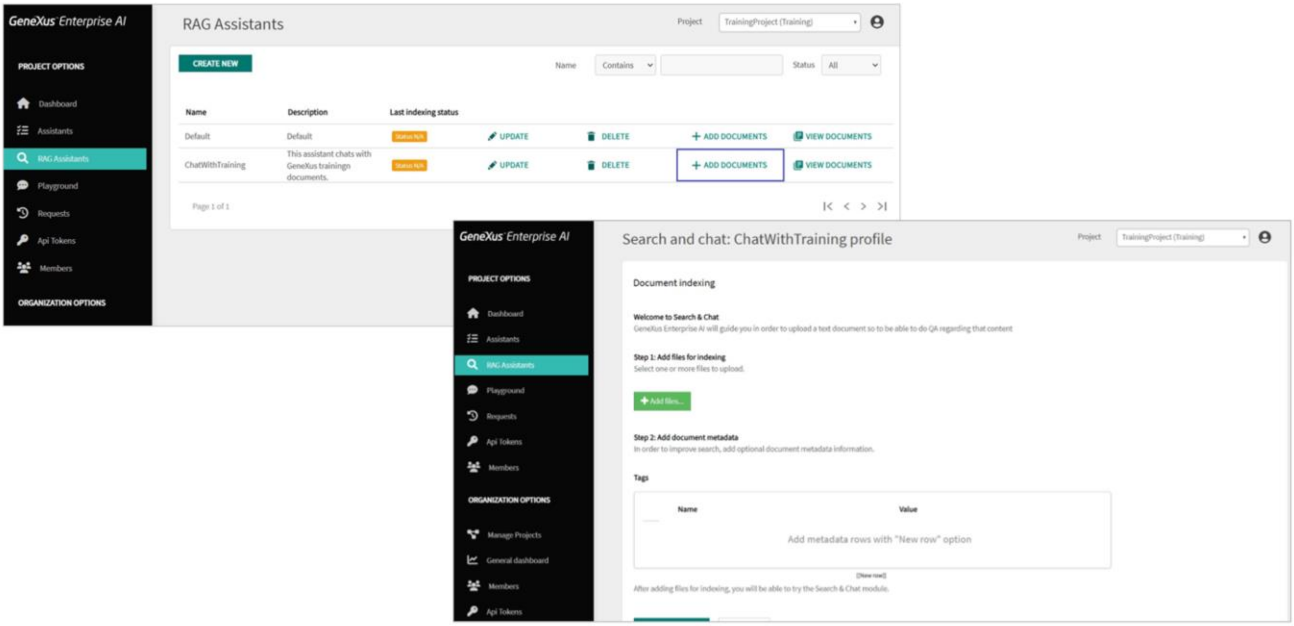
Profile Metadata: {}

General Information, Prompt, containing instructions to guide the assistant on how to approach and answer questions. These instructions establish clear guidelines for the assistant to provide relevant and useful answers based on the context provided.
This option indicates the number of fragments that are retrieved to expand the context.

Then this Message History option sets the number of historical messages that are taken into account in the conversation. It is useful for tracking the interaction history and understanding the context gathered in the conversation.

If this value is set to 4, it means that the last 4 interactions will be considered. The minimum value it can take is 0, which indicates that the conversation history is of no interest. When the value is greater than 0, it is used together with the message indicated in the following History Prompt option.

RAG Assistants



GeneXus Enterprise AI

RAG Assistants

Project: TrainingProject (Training)

Name: Contains
Status: All

Name	Description	Last indexing status				
Default	Default	Status N/A	UPDATE	DELETE	+ ADD DOCUMENTS	VIEW DOCUMENTS
ChatWithTraining	This assistant chats with GeneXus trainingn documents.	Status N/A	UPDATE	DELETE	+ ADD DOCUMENTS	VIEW DOCUMENTS

Page 1 of 1

GeneXus Enterprise AI

Search and chat: ChatWithTraining profile

Project: TrainingProject (Training)

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 QA 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.

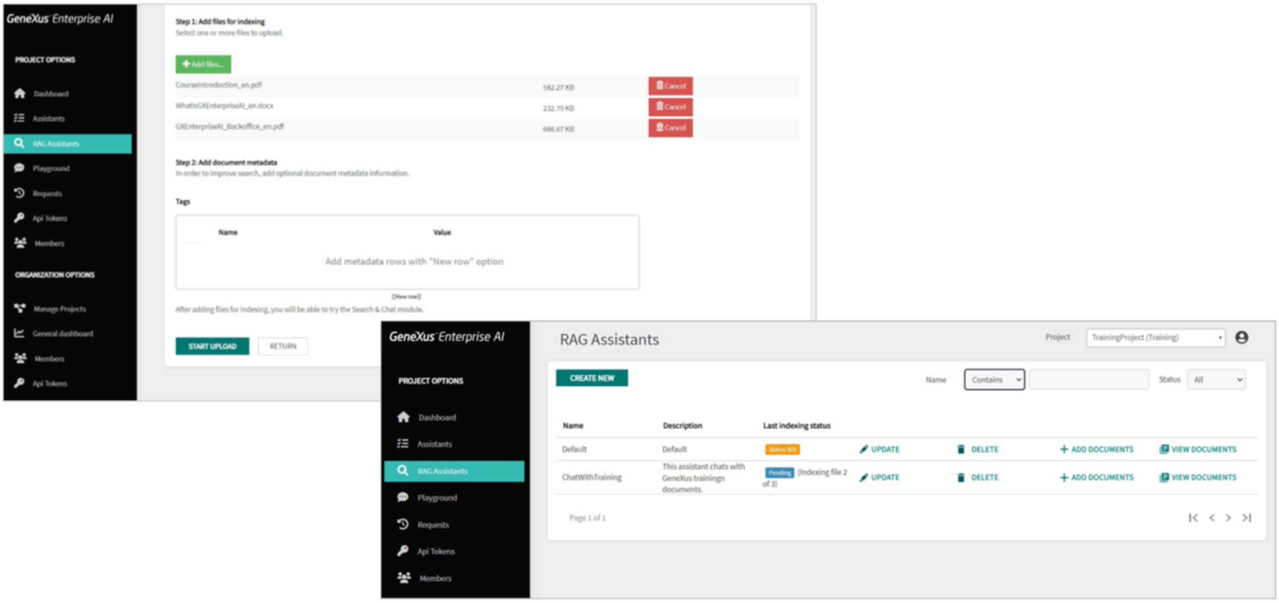
Next, we see the options for configuring the model used by the assistant to generate the answer. This includes the service provider, model name, temperature, maximum token limit, and other parameters that affect how answers are generated.

Lastly, the Retrieval tab specifies how the information is retrieved. We leave the default values.
OK. Now let's load the documents. We click on Add Documents. The Add Files button allows loading files of various formats:. .txt, .pdf, .docx, .pptx, .xlsx, .odt, .odp, .ods, .xlsx, .epub, .json, .jsonl, and .csv.

In our example, we are going to load a small set of PDF and Word documents corresponding to GeneXus Training material. We click on Add Files.

GeneXus by Globant

RAG Assistants



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

+ Add files...

CourseIntroduction_en.pdf	582.27 KB	Cancel
WhatisGXEnterpriseAI_en.docx	232.79 KB	Cancel
GXEnterpriseAI_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

[[New row]]

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

START UPLOAD RETURN

RAG Assistants

Project TrainingProject (Training)

CREATE NEW

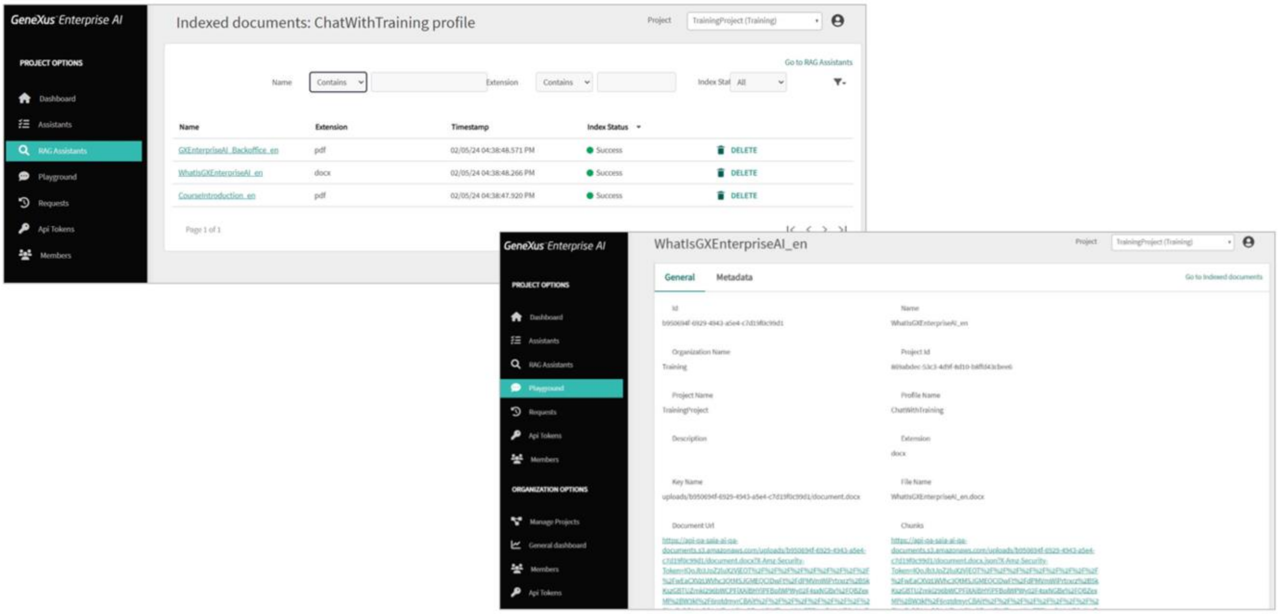
Name Contains Status All

Name	Description	Last indexing status				
Default	Default	Status N/A	UPDATE	DELETE	+ ADD DOCUMENTS	VIEW DOCUMENTS
ChatWithTraining	This assistant chats with GeneXus trainingn documents.	Pending (Indexing file 2 of 3)	UPDATE	DELETE	+ ADD DOCUMENTS	VIEW DOCUMENTS

Page 1 of 1

Once the documents have been added, we click on Start upload. To view the uploaded documents, each with its details, we click on View Documents.

RAG Assistants



Indexed documents: ChatWithTraining profile

Name: Contains — Extension: Contains — Index Stat: All

Name	Extension	Timestamp	Index Status	
GXEnterpriseAI_Backoffice_en	pdf	02/05/24 04:38:48.571 PM	● Success	DELETE
WhatIsGXEnterpriseAI_en	docx	02/05/24 04:38:48.266 PM	● Success	DELETE
CourseIntroduction_en	pdf	02/05/24 04:38:47.920 PM	● Success	DELETE

WhatIsGXEnterpriseAI_en

Id b95069df-6929-4943-a5e4-c7d19f0c99d1	Name WhatIsGXEnterpriseAI_en
Organization Name Training	Project Id 80abdec-53c3-4d9f-8d10-b8ffd43cbee6
Project Name TrainingProject	Profile Name ChatWithTraining
Description	Extension docx
Key Name uploads/b95069df-6929-4943-a5e4-c7d19f0c99d1/document.docx	File Name WhatIsGXEnterpriseAI_en.docx
Document Url https://api-qa-saia-ai-qa-documents.s3.amazonaws.com/uploads/b95069df-6929-4943-a5e4-c7d19f0c99d1/document.docx?X-Amz-Security-Token=IQo.Jb3JpZ2luX2VjEOT%2F%2F%2F%2F%2F%2F%2F%2F%2F...	Chunks https://api-qa-saia-ai-qa-documents.s3.amazonaws.com/uploads/b95069df-6929-4943-a5e4-c7d19f0c99d1/document.docx.json?X-Amz-Security-Token=IQo.Jb3JpZ2luX2VjEOT%2F%2F%2F%2F%2F%2F%2F%2F%2F...

By selecting the name of the file we see all the associated information, and we can view and download the file from the URL. Good. We have already created our RAG Assistant and uploaded the corresponding files. We are now ready to test it.

We will do it next from the Playground option of the menu.

GeneXus by Globant

GeneXus™ by Globant

training.genexus.com