

Converse com documentos – RAG Assistant



Alejandra Caggiano

Vimos que GeneXus Enterprise AI permite criar diferentes tipos de assistentes de inteligência artificial. Em particular, já conseguimos criar uma conversa interativa.

Queremos agora poder definir um assistente que nos permita conversar com documentos. Para isso, vamos trabalhar com assistentes RAG.

Retrieval Augmented Generation (RAG)

A Geração Aumentada de Recuperação (RAG) é uma abordagem que combina a recuperação de informação a partir de dados não estruturados e a geração de texto para melhorar o desempenho em tarefas como pode ser a resposta a perguntas.

Retrieval Augmented Generation (RAG)

- Ingestão de dados
- Recuperação
- Geração
- Interação com o usuário final

Este processo é composto pelas quatro fases seguintes:

- A primeira fase é a **entrada de dados**: envolve a carga de vários tipos de documentos, em diferentes formatos e a partir de múltiplas fontes.
- Segue-se então a fase de **Recuperação**: Nesta etapa inicia-se o processo de recuperação de dados, aproveitando a informação previamente carregada e organizada. É realizada uma busca seletiva sobre um conjunto de documentos, identificando a informação relacionada e reduzindo eficientemente o espaço de busca. Essa abordagem garante que a atenção se concentre na informação mais relevante e significativa.
- A próxima fase é a de **Geração**: O foco aqui está na geração de respostas relevantes e contextualmente consistentes. Neste processo, o sistema utiliza a configuração do assistente RAG para saber qual modelo acessar e com quais parâmetros. Este assistente incorpora os elementos necessários para definir a estratégia de busca e conseguir coerência e relevância no contexto gerado.
- A fase final é a **Interação com o usuário final**: GeneXus Enterprise AI facilita uma comunicação fluida e eficiente entre os usuários finais e os assistentes RAG, completando o ciclo e fornecendo respostas às consultas de forma eficiente.

Retrieval Augmented Generation (RAG)

Exemplo:

Converse com documentos de GeneXus Training

Bom. Como propusemos no início, nosso objetivo agora é criar um assistente que nos permita conversar com um conjunto de documentos, e faremos isso com documentos de GeneXus Training.

RAG Assistants

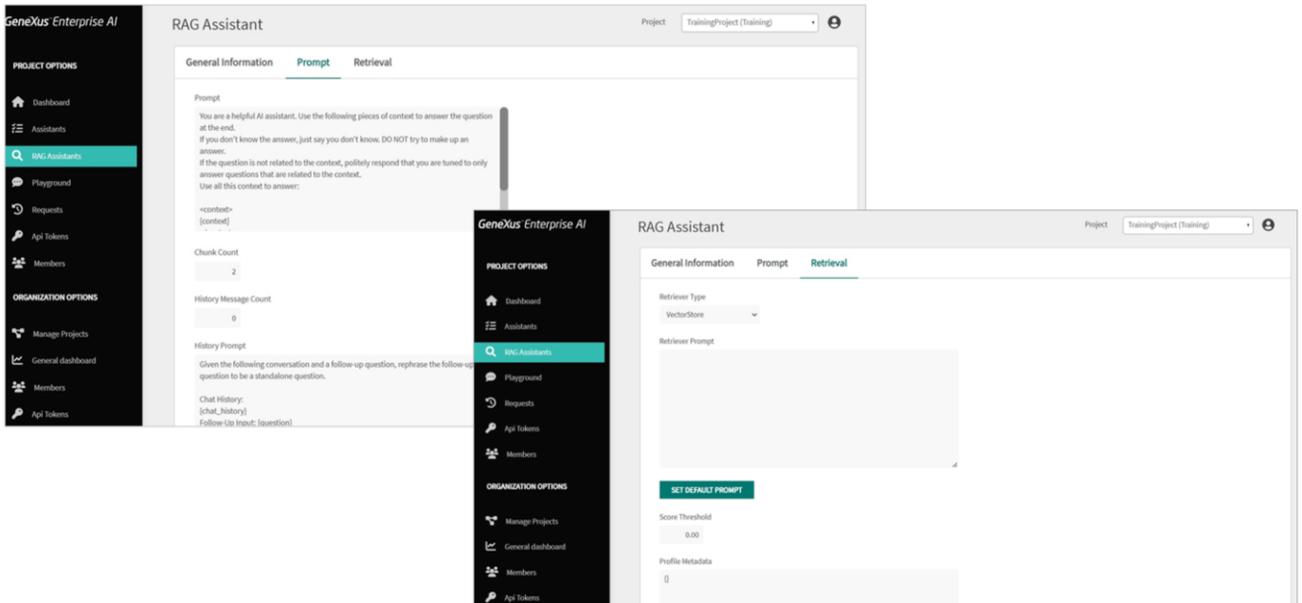
The image displays two screenshots of the GeneXus Enterprise AI interface. The top screenshot shows the 'RAG Assistants' management page for a project named 'TrainingProject (Training)'. It features a 'CREATE NEW' button and a table with columns for Name, Description, and Last indexing status. The table contains one entry: 'Default' with a 'PENDING' status. Below the table are navigation controls for 'Page 1 of 1'. The bottom screenshot shows the configuration form for a new RAG Assistant. The 'Name' field is 'ChatWithTraining' and the 'Description' field contains 'This assistant chats with GeneXus trainings documents'. The 'Status' is set to 'Enabled'. Under 'Embeddings Settings', the 'Provider Name' is 'openai', and there are fields for 'Model Name' and 'apiKey'.

Então entramos na plataforma, selecionamos o projeto sobre o qual vamos trabalhar, e no menu escolhemos RAG Assistant. Por padrão, ao abrir esta seção vemos um RAG Assistant chamado Default, que pode ser personalizado ou podemos criar novos.

Pressionamos Create new. Colocamos como nome ChatWithGXTraining e inserimos uma breve descrição

Pressionamos Confirm.

RAG Assistants



Na opção Update podemos personalizar a definição do assistente, conforme seja necessário. O conjunto de settings está organizado nestas guias:

A Informação geral, o Prompt, que contém instruções que orientam o assistente sobre como abordar e responder perguntas. Estas instruções estabelecem diretrizes claras para que o assistente forneça respostas relevantes e úteis baseadas no contexto fornecido.

Esta opção indica o número de fragmentos que são recuperados para aumentar o contexto.

Depois, esta opção de Histórico de mensagens estabelece o número de mensagens históricas que são levadas em consideração na conversa. Isso é útil para rastrear o histórico de interação e compreender o contexto compilado na conversa.

Se este valor for estabelecido em 4, significa que estamos interessados em considerar as últimas 4 interações:

O valor mínimo que pode assumir é 0, o que indica que o histórico de conversas não é de interesse. Quando o valor for maior que 0, é utilizado em conjunto com a mensagem indicada na seguinte opção History Prompt.

Vemos então as opções para estabelecer a configuração do modelo utilizado

pelo assistente para gerar a resposta. Isto inclui o provedor de serviços, o nome do modelo, a temperatura, o limite máximo de tokens e outros parâmetros que afetam a forma como são geradas as respostas.

Finalmente, a aba de Recuperação, que especifica como é recuperada a informação.

Deixamos os valores padrão.

RAG Assistants

The image shows two overlapping screenshots of the GeneXus Enterprise AI interface. The top screenshot displays the 'RAG Assistants' management page for a project named 'TrainingProject (Training)'. It features a 'CREATE NEW' button and a table listing existing assistants. The 'ChatWithTraining' assistant is highlighted, with its 'Last indexing status' set to 'Indexing' and a '+ ADD DOCUMENTS' button circled in blue. The bottom screenshot shows the 'Search and chat: ChatWithTraining profile' page, which guides the user through document indexing. It includes a 'Welcome to Search & Chat' message, instructions on how to upload documents, and a 'Step 1: Add files for indexing' section with an '+ Add Files...' button. A 'Step 2: Add document metadata' section is also visible, featuring a table for adding metadata rows.

Name	Description	Last indexing status	UPDATE	DELETE	+ ADD DOCUMENTS	VIEW DOCUMENTS
Default	Default	Indexing	[UPDATE]	[DELETE]	[+ ADD DOCUMENTS]	[VIEW DOCUMENTS]
ChatWithTraining	This assistant chats with GeneXus trainings documents.	Indexing	[UPDATE]	[DELETE]	[+ ADD DOCUMENTS]	[VIEW DOCUMENTS]

Name	Value
Add metadata rows with "New row" option	

Bem. Agora vamos fazer a carga dos documentos. Pressionamos Add Documents.

O botão Add Files permite realizar a carga de arquivos de diversos formatos:

.txt, .pdf, .docx, .pptx, .xlsx, .odt, .odp, .ods, .xlsx, .epub, .json, .jsonl e .csv. .

RAG Assistants

The screenshot displays the GeneXus Enterprise AI RAG Assistants interface. On the left is a navigation sidebar with options like Dashboard, Assistants, RAG Assistants (selected), Playground, Requests, Api Tokens, and Members. The main content area is divided into two steps:

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

File Name	Size	Action
CourseIntroduction_en.pdf	582.27 KB	Cancel
WhatsGXEnterpriseAI_en.docx	232.79 KB	Cancel
GXEnterpriseAI_Backoffice_en.pdf	696.67 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]

The bottom part of the screenshot shows the 'RAG Assistants' management view for the 'TrainingProject (Training)' project. It includes a 'CREATE NEW' button and a table of existing assistants:

Name	Description	Last indexing status	Actions
Default	Default	Indexing	[UPDATE] [DELETE] [+ ADD DOCUMENTS] [VIEW DOCUMENTS]
ChatWithTraining	This assistant chats with GeneXus training documents.	Indexing (Indexing file 2 of 3)	[UPDATE] [DELETE] [+ ADD DOCUMENTS] [VIEW DOCUMENTS]

Page 1 of 1

Em nosso exemplo, vamos carregar um pequeno conjunto de pdfs e docs que correspondem ao material de GeneXus training.

Pressionamos Add Files

Uma vez carregados os documentos, pressionamos Start upload

RAG Assistants

The screenshot displays the GeneXus Enterprise AI interface. On the left is a navigation sidebar with options: Dashboard, Assistants, RAG Assistants (selected), Playground, Requests, API Tokens, and Members. The main area is titled 'Indexed documents: ChatWithTraining profile' and shows a table of indexed documents. Below the table, the 'View Documents' modal is open, showing metadata for the document 'WhatsGXEnterpriseAI_en.docx'.

Name	Extension	Timestamp	Index Status	
GXEnterpriseAI_Backoffice_en	pdf	02/05/24 04:38:48.571 PM	Success	DELETE
WhatsGXEnterpriseAI_en	docx	02/05/24 04:38:48.266 PM	Success	DELETE
CourseProduction_en	pdf	02/05/24 04:38:47.520 PM	Success	DELETE

General	Metadata
ID	Name
0000004f-6029-4943-a564-c7028f6c9192	WhatsGXEnterpriseAI_en
Organization Name	Project ID
Training	0000000c-53c3-4d5f-8d10-baff6f6c3b66
Project Name	Profile Name
TrainingProject	ChatWithTraining
Description	Extension
	docx
Key Name	File Name
uploads/0000004f-6029-4943-a564-c7028f6c9192/document.docx	WhatsGXEnterpriseAI_en.docx
Document Url	Chunks
https://api-gx.globant.com/documents/0000000c-53c3-4d5f-8d10-baff6f6c3b66/documents/0000004f-6029-4943-a564-c7028f6c9192/document.docx?Access-Token=0000000c-53c3-4d5f-8d10-baff6f6c3b66	https://api-gx.globant.com/documents/0000000c-53c3-4d5f-8d10-baff6f6c3b66/documents/0000004f-6029-4943-a564-c7028f6c9192/document.docx?Access-Token=0000000c-53c3-4d5f-8d10-baff6f6c3b66

Para ver os documentos carregados, cada um com seu detalhe, pressionamos View Documents

Selecionando o nome do arquivo vemos toda a informação associada, podendo visualizar e baixar o arquivo a partir da URL.

Bom. Já criamos nosso RAG Assistant e o carregamos com os arquivos correspondentes. Estamos então em condições de testá-lo.

Faremos isso a seguir a partir da opção Playground do menu.

GeneXus[™]
by **Globant**

training.genexus.com