GeneXus by Globant

Converse com documentos – RAG Assistant

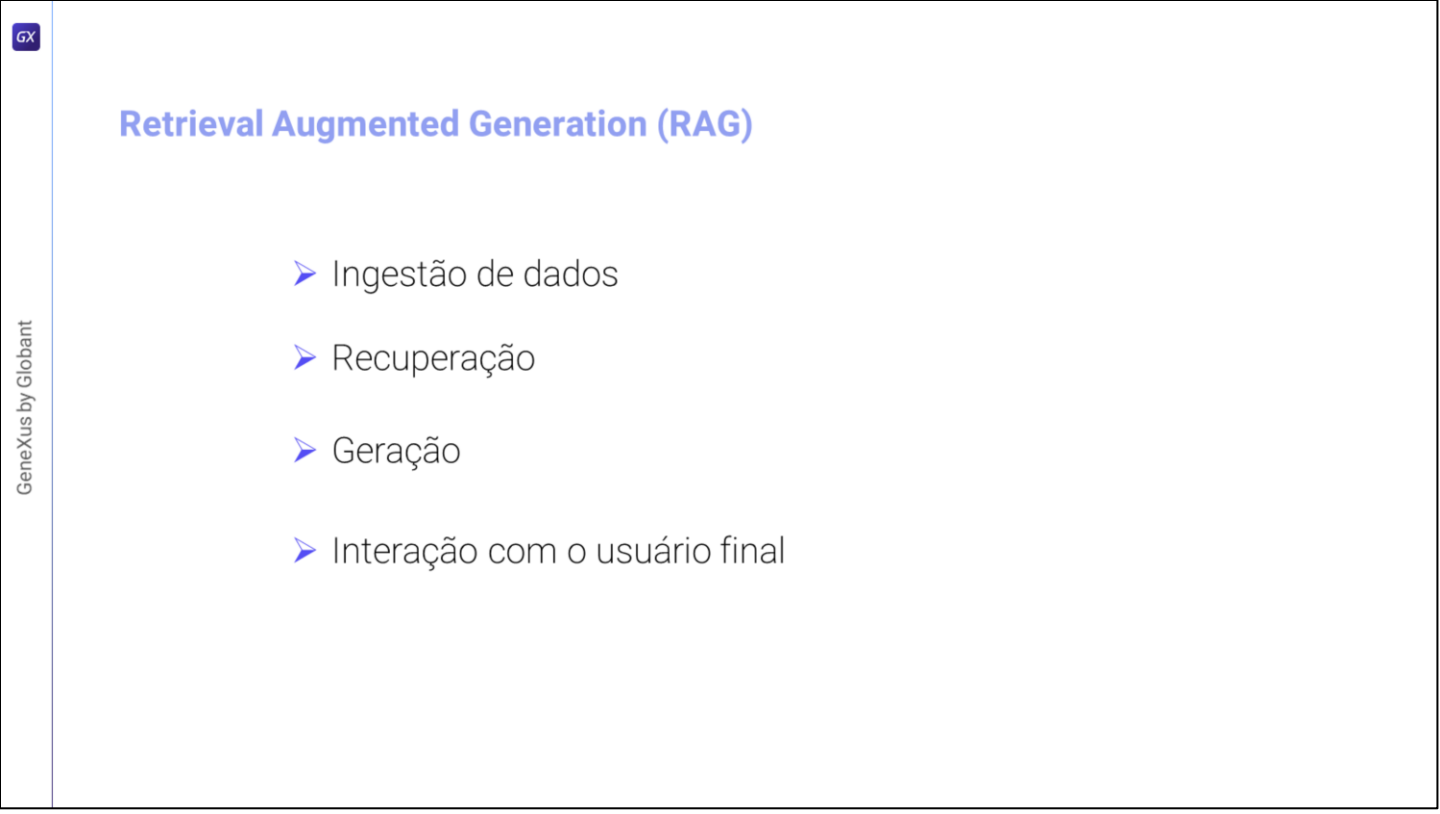
Alejandra Caggiano

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.



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.

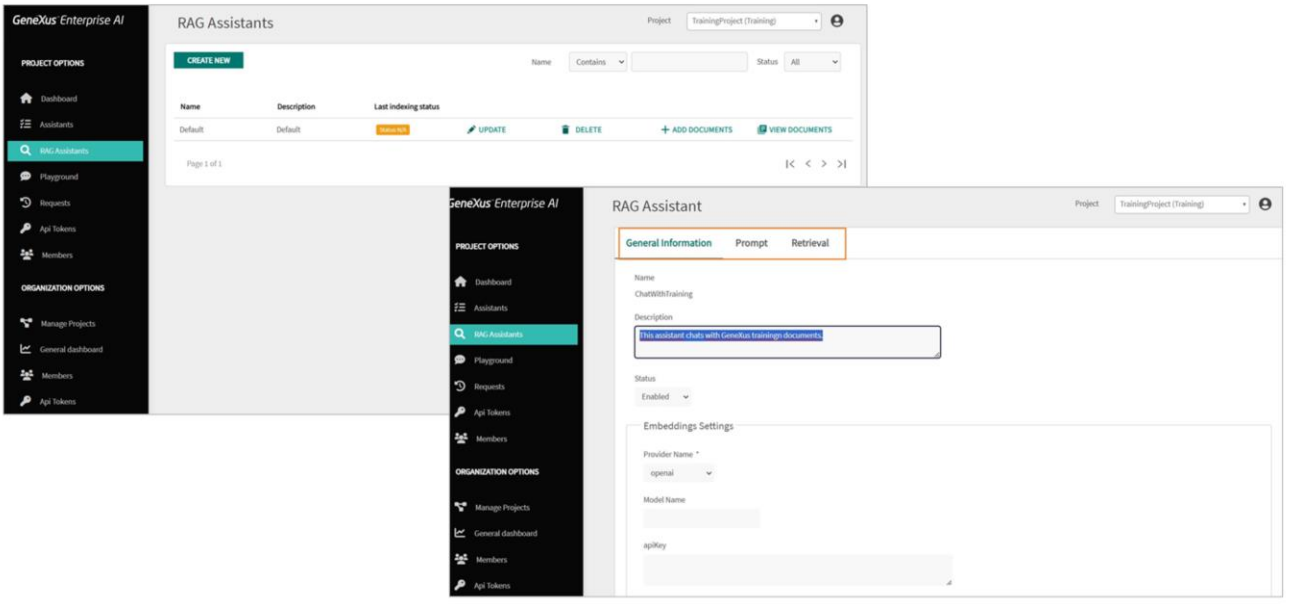
Retrieval Augmented Generation (RAG)

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

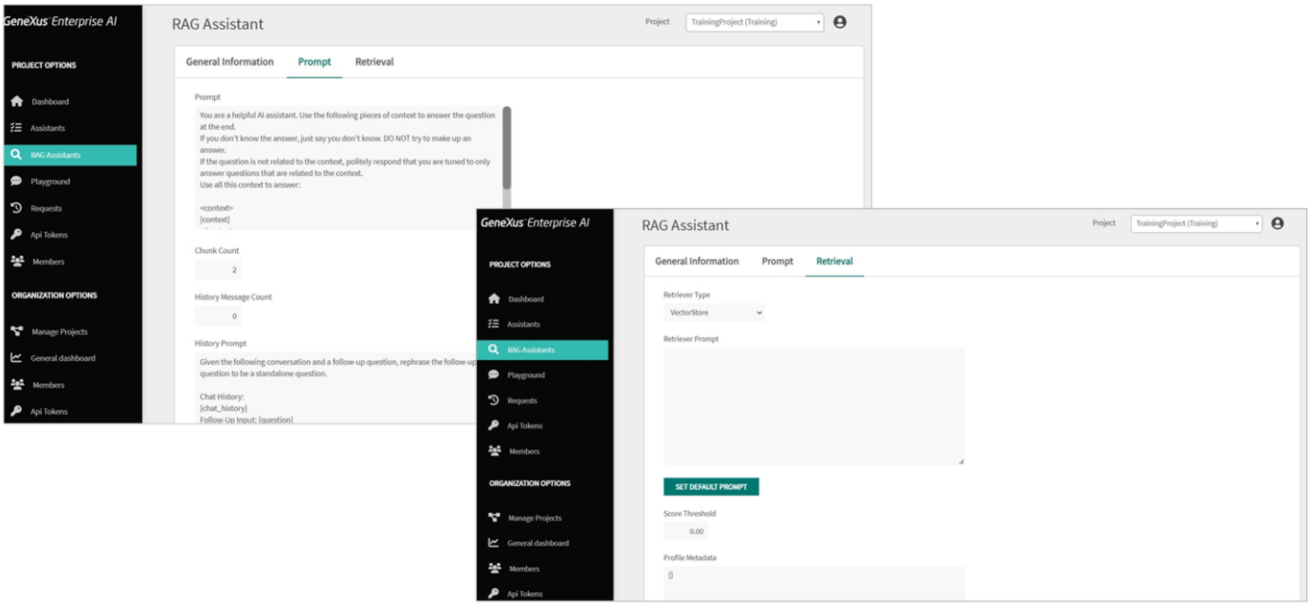


Então entramos na plataforma, selecionamos o projeto sobre o qual vamos trabalhar, e no menu escolhemos RAG Assistant. Por padrão, ao abrir esta seção vemos um RAG Assistant chamado Default, que pode ser personalizado ou podemos criar novos.

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.

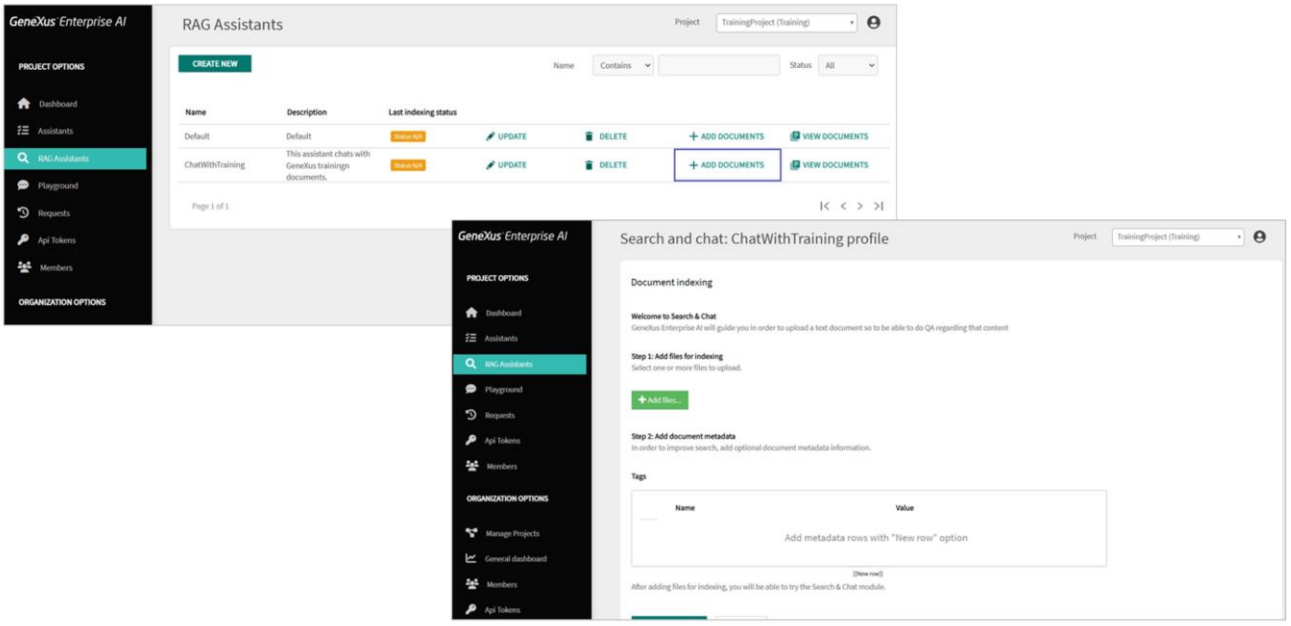
Vemos então as opções para estabelecer a configuração do modelo utilizado

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

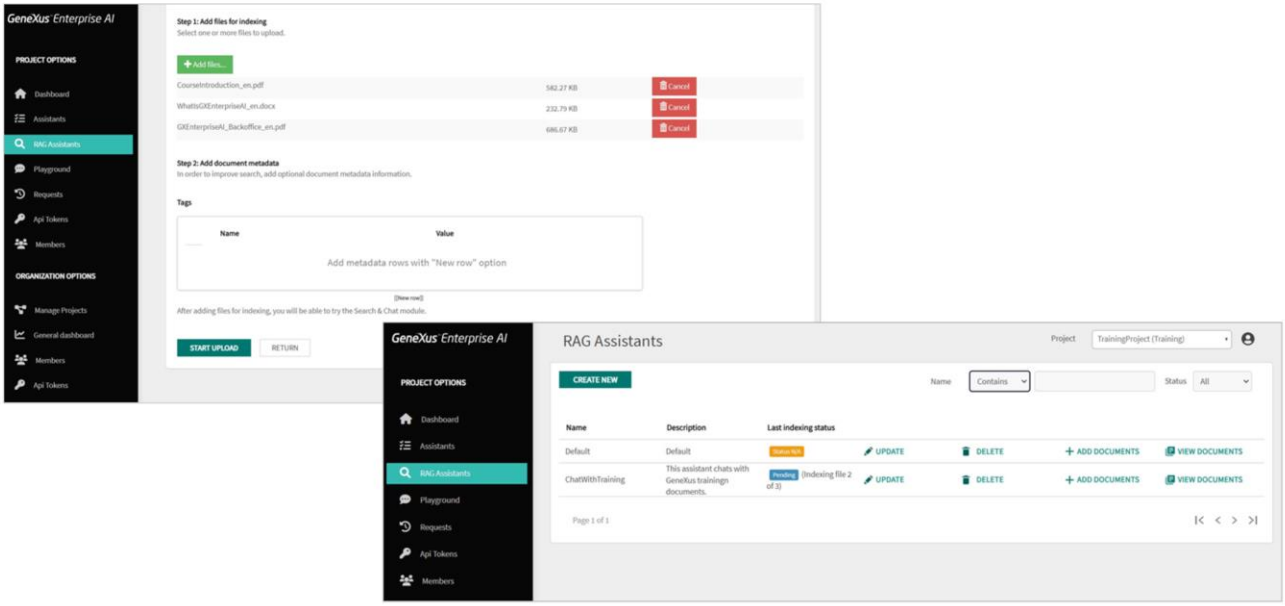


Bem. Agora vamos fazer a carga dos documentos. Pressionamos Add Documents.

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

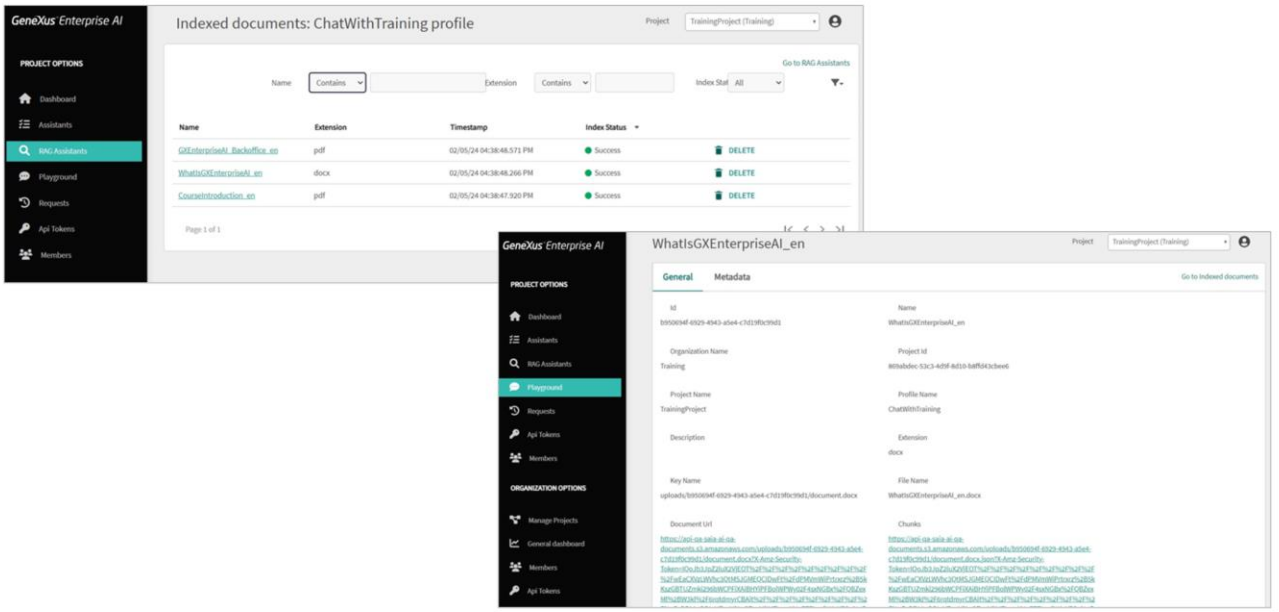


Em nosso exemplo, vamos carregar um pequeno conjunto de pdfs e docs que correspondem ao material de GeneXus training.

Pressionamos Add Files

Uma vez carregados os documentos, pressionamos Start upload

RAG Assistants



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

GeneXus™
by Globant

training.genexus.com