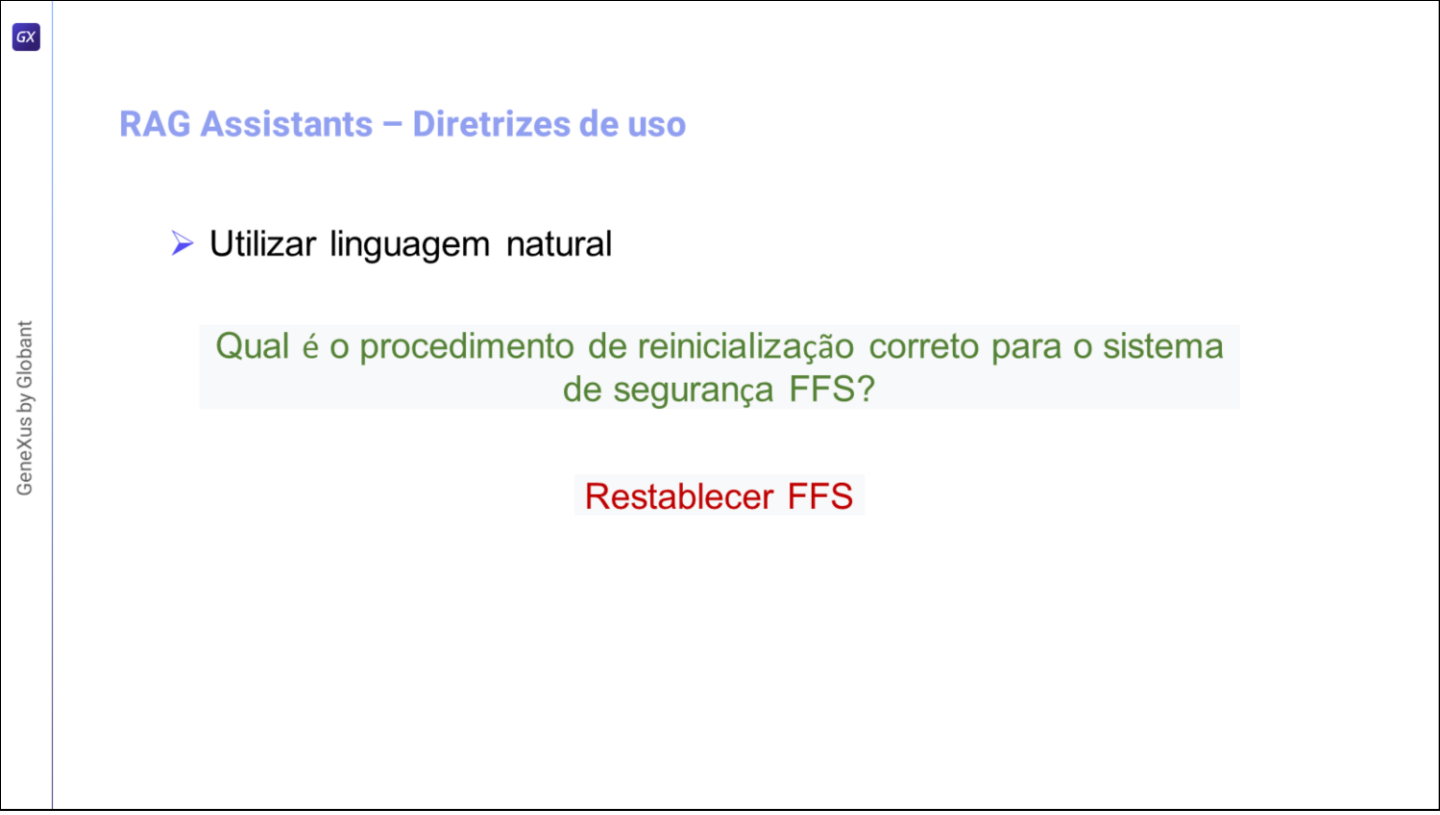
GeneXus by Globant

RAG Assistants – Diretrizes de uso e Configuração

Alejandra Caggiano

Avançando um pouco mais, agora que já sabemos como criar e testar um assistente, vejamos algumas diretrizes sobre como fazer perguntas corretas para obter informação que fez parte da entrada de documentos de um determinado Assistente RAG.

Vejamos então algumas considerações gerais:

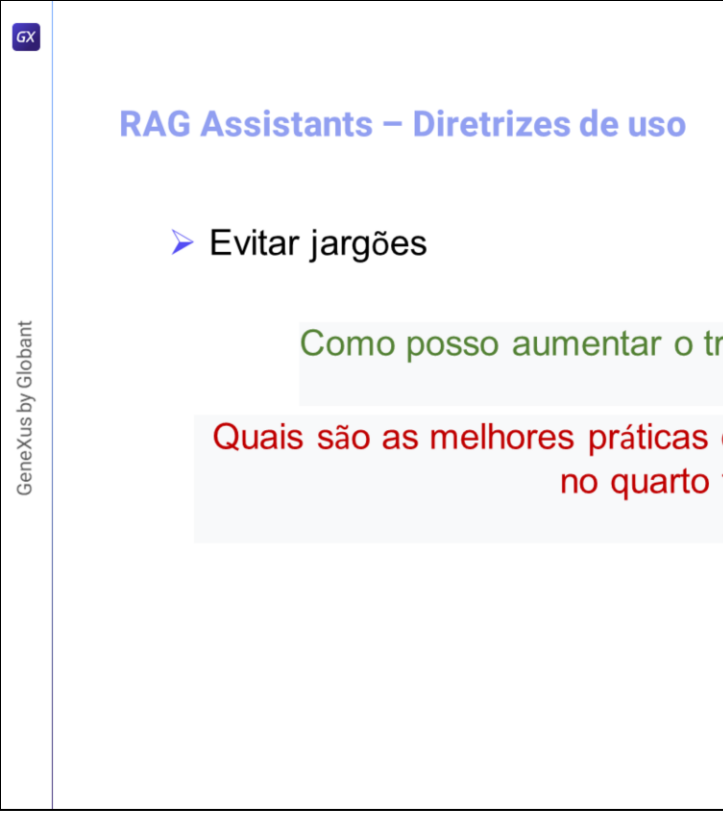


A primeira consideração é **utilizar linguagem natural**: Isso significa escrever perguntas com naturalidade, como se você estivesse falando com uma pessoa.

Não importa se for cometido algum erro ortográfico ou se são utilizados sinônimos. O importante é transmitir corretamente a intenção.

Para dar um exemplo, uma pergunta correta pode ser: Qual é o procedimento de reinicialização correto para o sistema de segurança FFS?

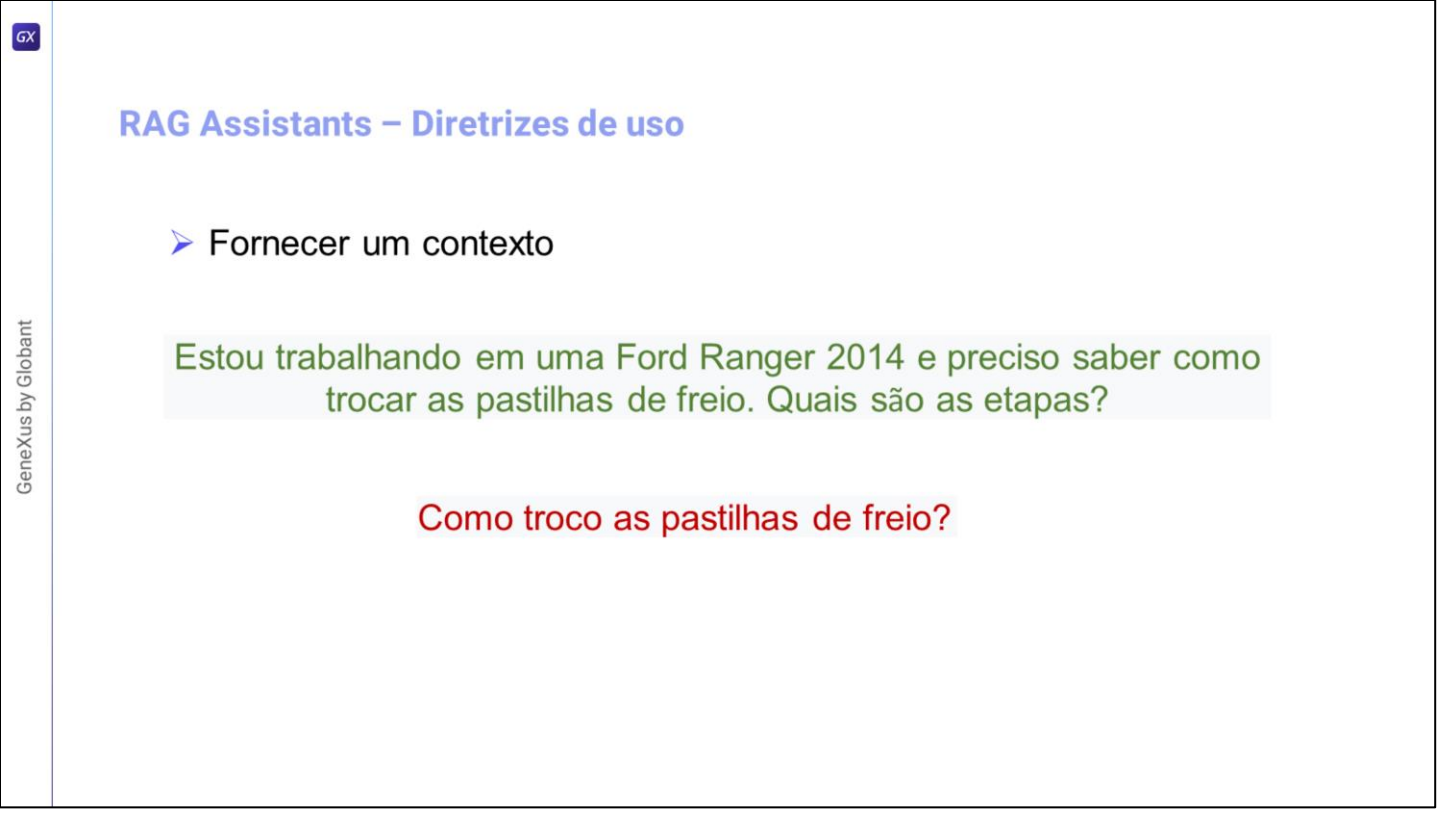
Enquanto uma expressão incorreta seria escrever simplesmente: Restabelecer FFS



Bem. Outra consideração a levar em conta é **evitar jargões**: Embora o assistente possa compreender termos técnicos, uma linguagem mais simples pode muitas vezes fornecer resultados mais precisos.

Como exemplo, uma expressão correta seria: Como posso aumentar o tráfego do meu site web?

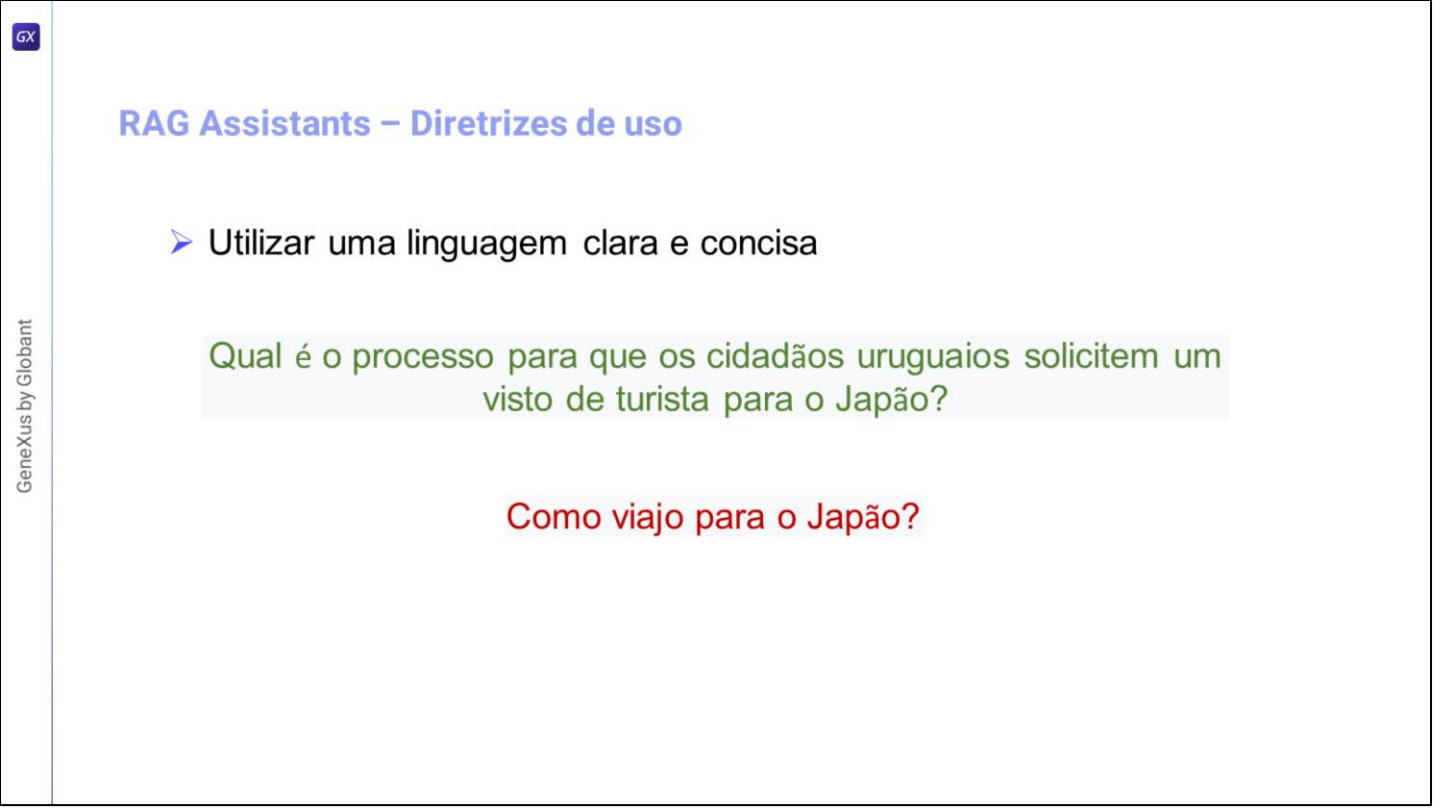
Enquanto uma expressão incorreta seria: Quais são as melhores práticas de SEO para dominar as SERP no quarto trimestre?



Outra consideração importante é a de **fornecer um contexto**: Adicionar um pouco de contexto ou histórico à consulta pode ajudar o assistente a compreender melhor o escopo da solicitação.

Uma expressão correta seria: Estou trabalhando em uma Ford Ranger 2014 e preciso saber como trocar as pastilhas de freio. Quais são as etapas?

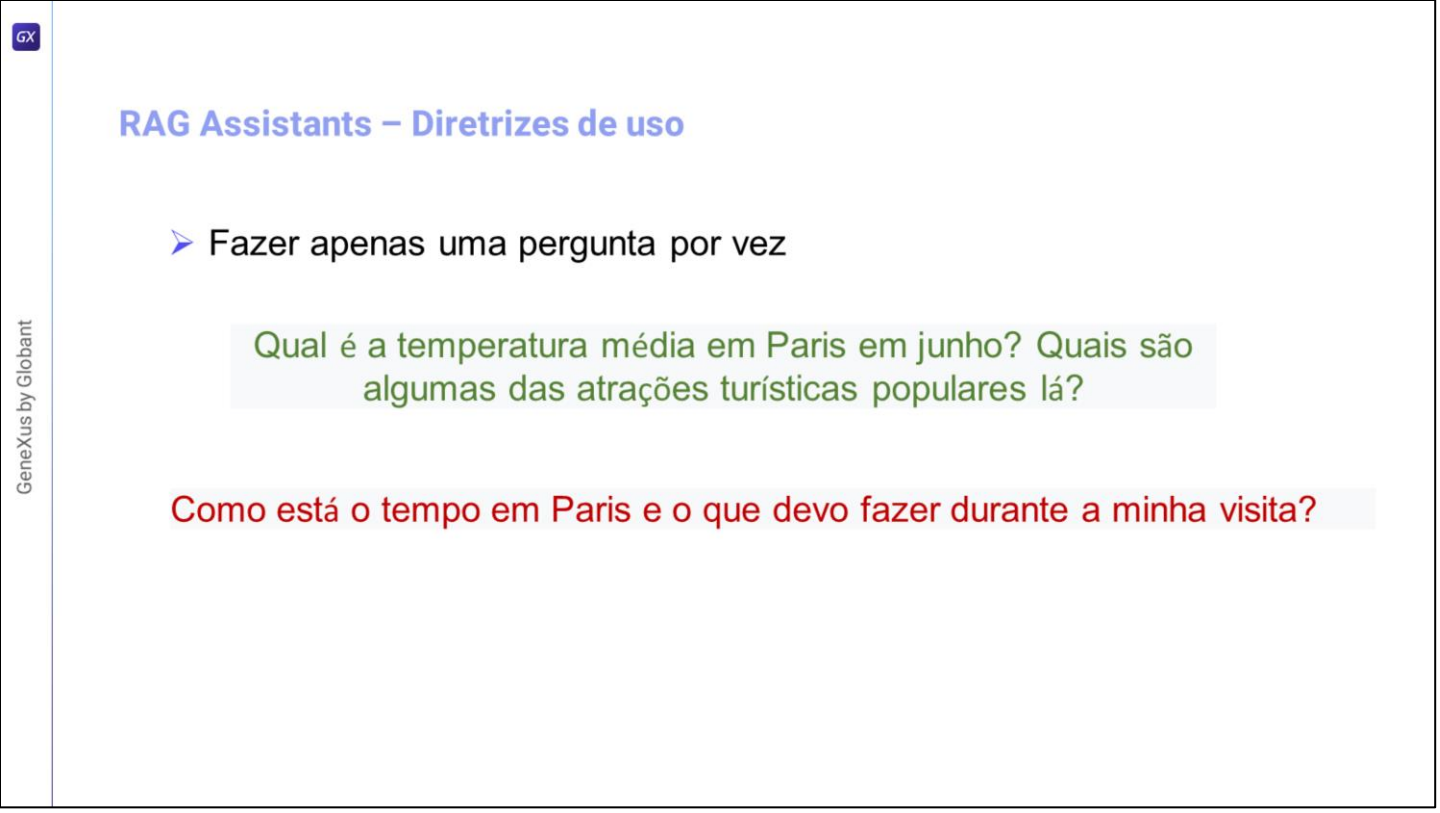
Enquanto não seria bom simplesmente perguntar: Como troco as pastilhas de freio?



Outra orientação recomendada é **utilizar uma linguagem clara e concisa**: É importante ser específico e articular com clareza o que se busca para receber informação mais precisa e relevante. Evitar a ambiguidade ajuda o assistente a fornecer respostas precisas.

Um exemplo correto seria consultar: Qual é o processo para que os cidadãos uruguaios solicitem um visto de turista para o Japão?

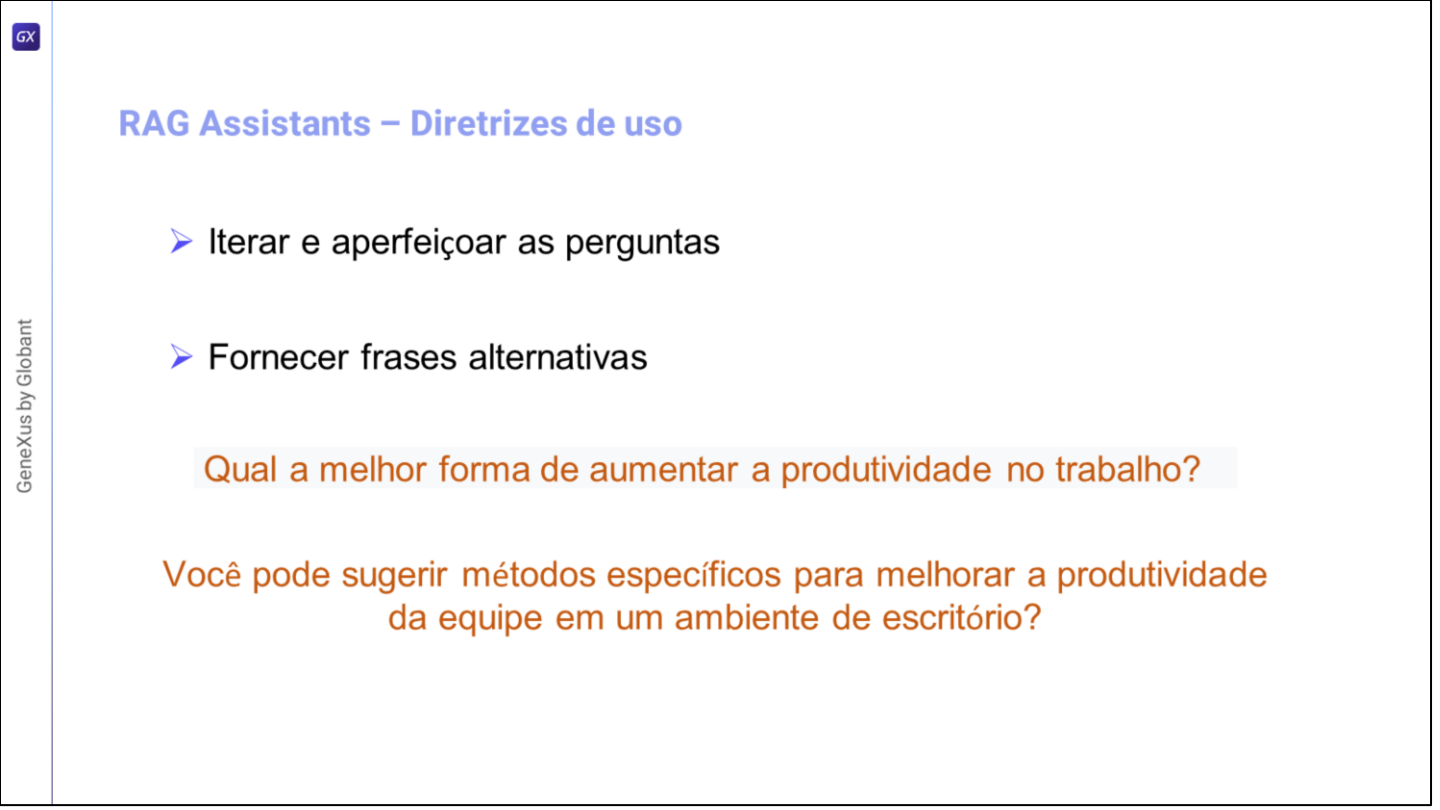
Por outro lado, não seria recomendável simplesmente consultar: Como viajo para o Japão?



Outra boa orientação é **fazer apenas uma pergunta por vez**: Se precisarmos fazer uma consulta complexa, é bom considerar dividi-la em partes mais simples.

Um exemplo correto seria: Qual é a temperatura média em Paris em junho? Quais são algumas das atrações turísticas populares lá?

Por outro lado, não seria o melhor perguntar: Como está o tempo em Paris e o que devo fazer durante a minha visita?



Outra boa consideração é **iterar e aperfeiçoar as perguntas**: Se a resposta inicial for diferente da necessária, é bom reconstruir a pergunta de acordo com a resposta fornecida pelo assistente.

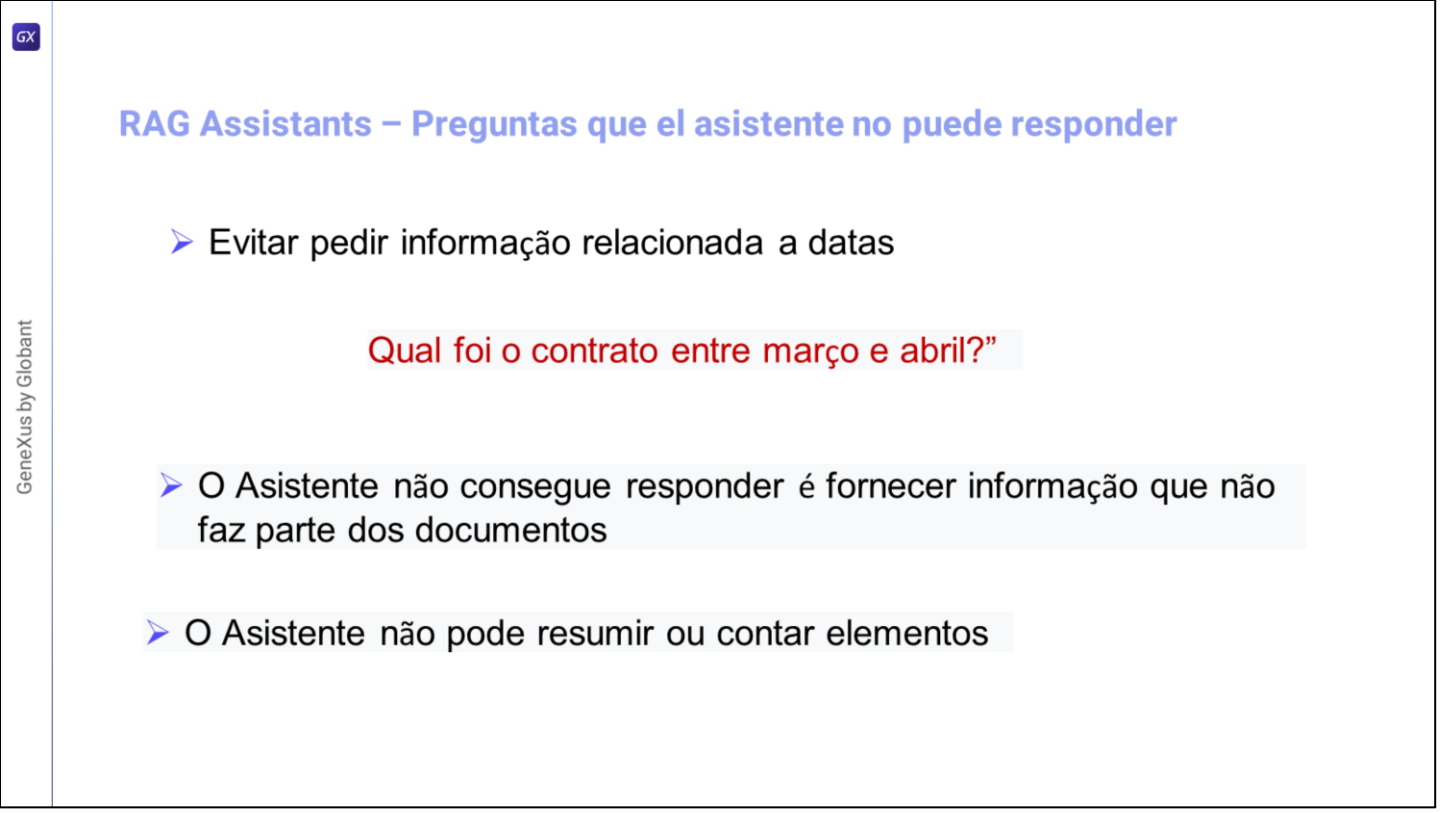
Também é bom **fornecer frases alternativas**: Isto significa testar com diferentes formas de formular uma consulta, incluindo sinônimos e variações, a fim de explorar a compreensão do assistente.

Vejamos como exemplo a seguinte consulta:

Qual a melhor forma de aumentar a produtividade no trabalho? Se o assistente não compreender a pergunta, ele poderá fornecer como resposta estratégias gerais sobre produtividade em vez de estratégias específicas no local de trabalho.

Então é possível reformular da seguinte forma:

Você pode sugerir métodos específicos para melhorar a produtividade da equipe em um ambiente de escritório? Essa reformulação, então, esclarece que o foco está concentrado na produtividade da equipe em um ambiente profissional, e não em conselhos de produtividade individuais.



Bom. Assim como acabamos de ver orientações e considerações para um bom uso na comunicação com um assistente, é importante levar em conta também que existem certas **perguntas que o assistente não consegue responder**.

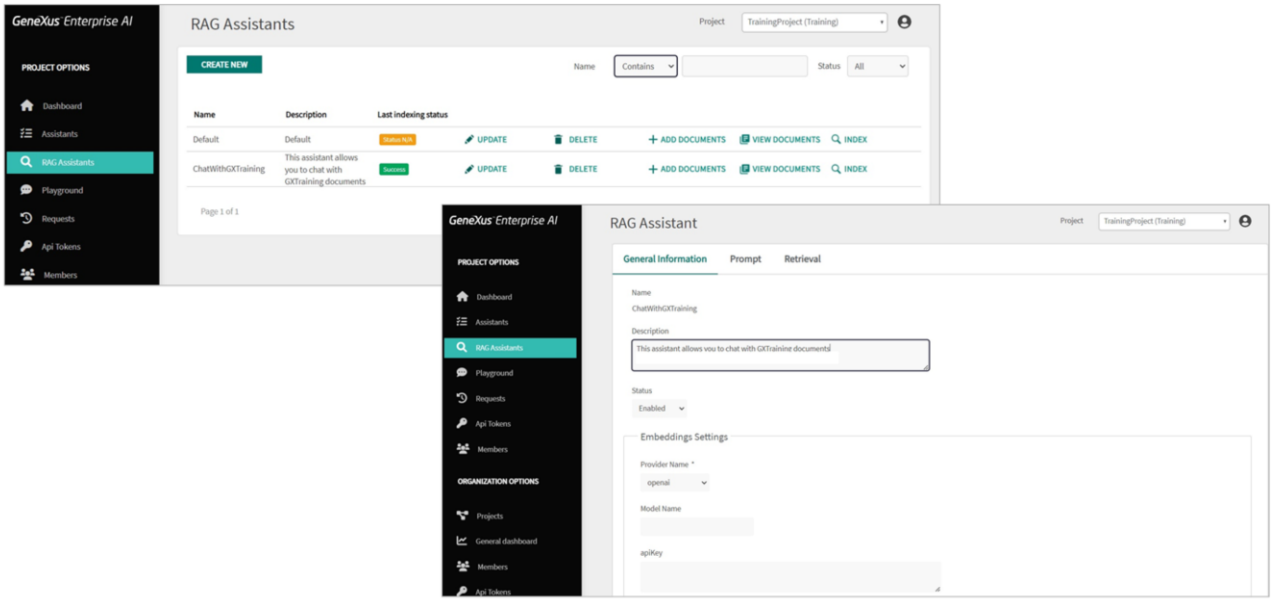
Por exemplo, é importante **evitar pedir informação relacionada a datas**. Algo deste estilo não seria recomendável: “Qual foi o contrato entre março e abril?”.

Não se espera que as comparações ou o contraste de informação produzam bons resultados;

Outra consulta que o assistente não consegue responder é fornecer informação que não faz parte dos documentos que foram utilizados para dar contexto ao Assistente RAG.

Ele não pode resumir ou contar elementos e não pode processar textos que aparecem em imagens.

RAG Assistants – Configuração



Bem. Agora que já conhecemos as considerações a levar em conta para ter uma boa comunicação com um assistente RAG, vamos aprofundar um pouco mais na sua configuração.

Como já vimos anteriormente, toda a interação com o componente de Busca e Chat é configurada através da seção RAG Assistants. É criado um assistente RAG padrão durante a inicialização e depois é possível modificar ou criar outros novos para alterar seu comportamento.

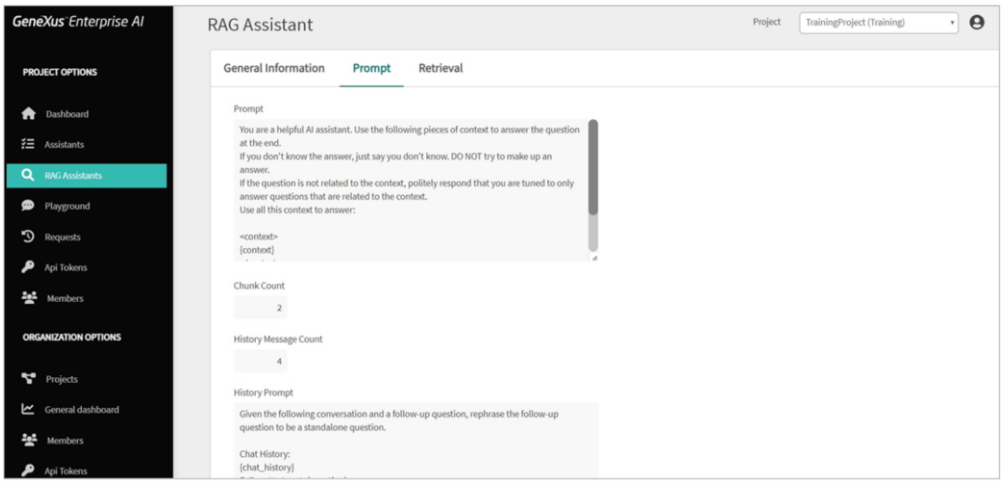
Uma vez criado, acessamos suas opções através do botão Update, e já sabemos que o conjunto dessas opções está organizado em três seções:

A seção **General Information**, onde são indicados detalhes que permitem compreender rapidamente as capacidades e características do assistente.

A seção **Prompt**: onde se indica informação sobre como está configurado o assistente para receber e processar consultas dos usuários.

E a seção **Retrieval**: onde se especifica como é recuperada a informação.

RAG Assistants – Configuração



Vamos para a seção Prompt

Como já sabemos, aqui podem ser configuradas 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. O valor padrão é este que estamos vendo.

Devem ser mantidas configuradas as variáveis de contexto e pergunta porque serão substituídas pela informação associada antes da interação.

Bem. A próxima opção define quantos fragmentos são recuperados para aumentar o contexto.

A opção Quantidade de mensagens do histórico estabelece a quantidade de mensagens do histórico que são consideradas na conversa.

Isso é útil para rastrear o histórico de interação e compreender o contexto coletado na conversa. Deve-se considerar que este valor se refere à pergunta do usuário final e à resposta associada. Ou seja, se estiver definido como 4, significa

que está interessado em considerar as últimas 4 interações:

• Pergunta anterior

• Resposta anterior

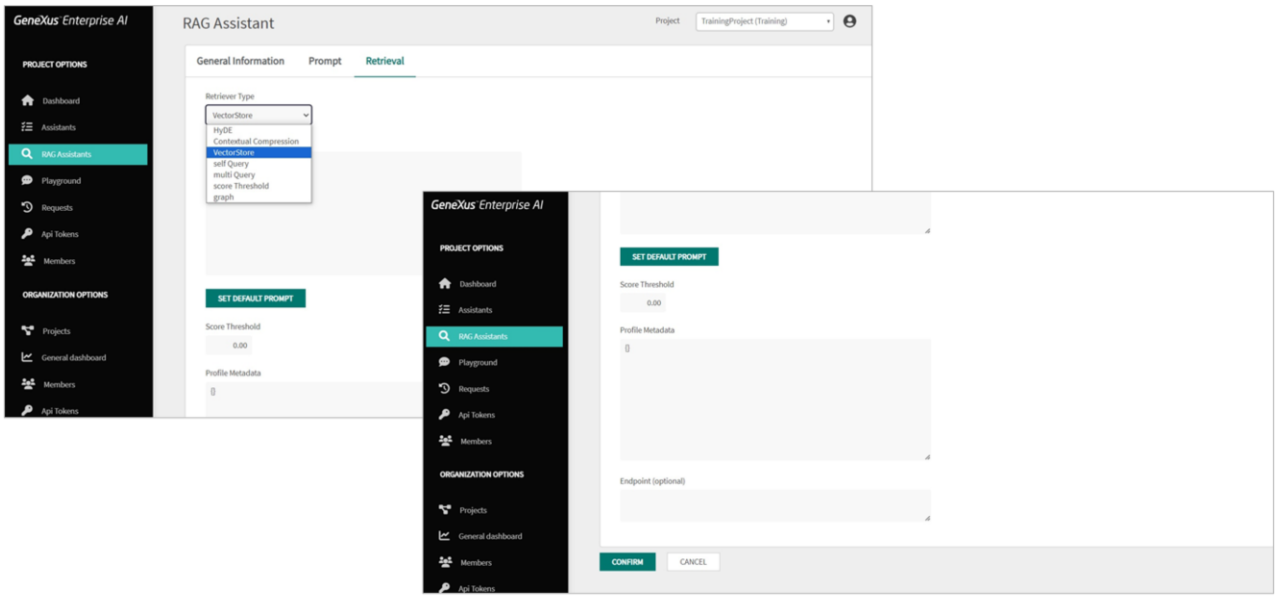
• Última pergunta

• e Última resposta

O valor mínimo que pode assumir é 0, o que indica que o histórico de conversas não é de interesse.

Quanto às opções de configuração do LLM abrangem a configuração do modelo utilizado pelo assistente para gerar a resposta, incluindo o provedor de serviços, o nome do modelo, a temperatura, o limite máximo de token e outros parâmetros que afetam a forma como são geradas as respostas.

RAG Assistants – Configuração



Passemos para a aba de **Recuperação**

Esta seção permite especificar como obter a informação enviada ao contexto.

Aqui é possível indicar o tipo de recuperador utilizado para obter informação. O valor padrão é VectorStore, que utiliza diretamente o VectorStore definido sem processamento prévio adicional.

Os outros valores que pode assumir são os seguintes:

Método de **incorporação de documentos hipotéticos**: Este método utiliza técnicas de incorporação para responder consultas, gerar respostas hipotéticas, incorporá-las no documento gerado e, em seguida, utilizá-lo como exemplo final.

Outro possível método de recuperação é a **Compressão Contextual**. Este método tenta melhorar as respostas retornadas pelas buscas de similaridade de documentos, levando em consideração de forma mais adequada o contexto da consulta.

Outra opção é a **Autoconsulta**. Nesse tipo de recuperação, primeiro consulta a si mesmo para recuperar informação de filtro baseada na consulta em linguagem natural. Em seguida, executa uma segunda consulta ao LLM com a

consulta e os filtros aplicados com base na primeira.

Também é possível escolher a **Consulta múltipla**. Este tipo de recuperador automatiza o processo de ajuste rápido através do uso de um LLM para gerar múltiplas consultas a partir de diferentes perspectivas para uma determinada consulta inicial do usuário. Para cada consulta, então, recupera um conjunto de documentos relevantes e toma a união única de todas as consultas para então obter um conjunto maior de documentos potencialmente relevantes.

Em seguida, a opção **Limite de pontuação**, utiliza o que é chamado de pesquisa recursiva de similaridade. Serão retornadas todas as correspondências de perguntas semelhantes com base no limite de pontuação mínimo indicado.

Por último, o tipo de recuperador **Gráfico** permite o uso de uma abordagem de representação de informação baseada em gráficos para sua recuperação.

E em seguida, nesta caixa é possível especificar a consulta que é enviada ao recuperador para buscar informação. Esta consulta pode ser uma pergunta ou uma solicitação específica.

Esta caixa Limite de pontuação define o valor mínimo válido para considerar a informação como válida quando for recuperada. Se não houver documentos válidos, não ocorre nenhuma interação com o LLM. O valor padrão é 0,0.

A caixa **metadata do perfil**. Pode conter metadata adicional relacionado ao perfil do recuperador. O valor padrão é um objeto vazio, mas em certos casos é conveniente poder ajustar cada acesso ao LLM para obter o resultado desejado.

Finalmente, na caixa **Endpoint** é possível indicar a URL que aponta para o servidor ou serviço específico onde estão hospedados os métodos ou modelos de recuperação.

GeneXus by Globant

GeneXus™
by Globant

training.genexus.com