O que é RAG e como essa técnica funciona

L
larissa-dubiella
11/12/2024

RAG (*Retrieval Augmented Generation*) é uma técnica utilizada para **ampliar a capacidade de resposta de LLMs, combinando o conhecimento interno do modelo de linguagem com sistemas de recuperação de informações**.

Ou seja, o modelo busca informações relevantes em bases de dados externas como bancos de dados ou documentos organizacionais antes de gerar uma resposta, permitindo acesso a dados atualizados, especializados ou muito específicos sem a necessidade de re-treinar o modelo.

Confira neste artigo:

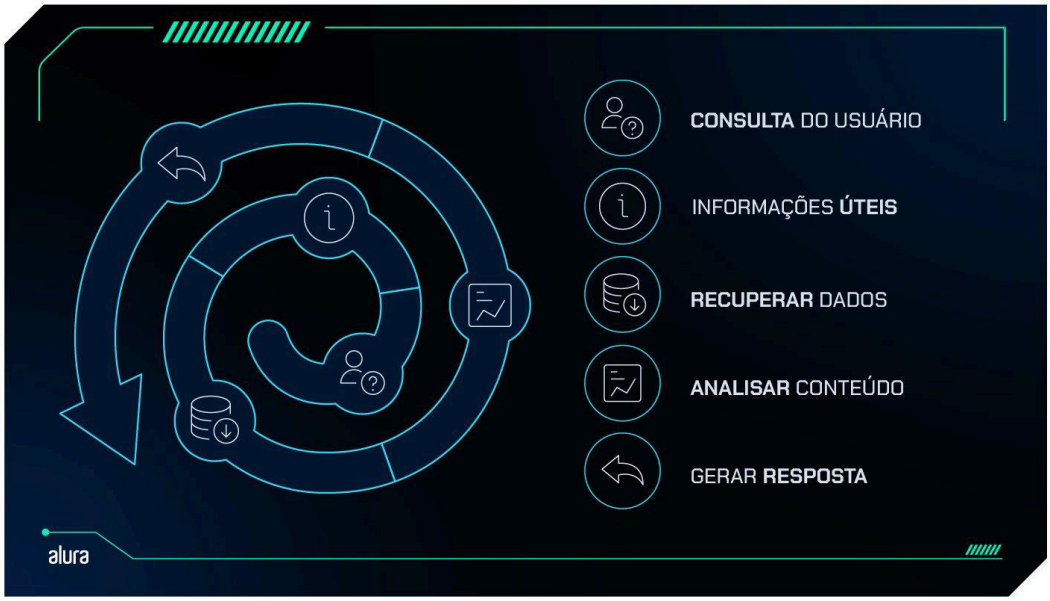
É uma ideia simples, mas poderosa, com aplicações práticas em qualquer área que se beneficie de um sistema com LLM.

O funcionamento do RAG pode ser entendido como uma parceria entre duas partes principais.

Primeiro, temos o *componente de recuperação*, que age buscando informações em fontes externas.

Em seguida, entra o *componente de geração*, onde o modelo de linguagem utiliza as informações recuperadas para criar respostas contextualizadas e de alta qualidade.

Dessa forma, temos uma solução para limitações dos LLMs em termos de precisão, relevância e controle de informações.



Por que RAG é importante?

Modelos de linguagem possuem limitações que comprometem sua confiabilidade, e o RAG pode ser a solução para diversas delas.

RAG é essencial para mitigar as seguintes limitações das LLMs:

- **Respostas Fabricadas**: Quando o modelo não possui uma resposta adequada, ele inventa informações - aquilo que chamamos de *alucinação*.
- **Desatualização de Dados**: Como os LLMs têm uma data de corte em relação ao treinamento, não conseguem acessar eventos ou mudanças recentes, o que reduz sua utilidade em diversos contextos.
- **Confusão Terminológica**: Termos semelhantes, mas com significados distintos em diferentes contextos, podem levar o modelo a gerar respostas imprecisas.
- **Falta de Transparência**: Os modelos não são capazes de citar fontes e indicar a origem do texto gerado.

Com a adoção de RAG, os LLMs se tornam ferramentas muito mais confiáveis, uma vez que são capazes de acessar e incorporar conhecimento externo.

Como o RAG funciona na prática?

O processo funciona assim: a pessoa usuária faz uma pergunta, o sistema identifica quais informações externas podem ser úteis, faz a busca e traz os dados para o modelo.

O modelo, então, analisa o conteúdo e produz uma resposta integrada, usando tanto o material recuperado quanto seu próprio repertório.

Esse fluxo garante que as respostas sejam mais precisas e atualizadas. Vamos pensar mais detalhadamente em cada parte desse processo.

O modelo se conecta com bases de conhecimento externas, como APIs, bancos de dados ou documentos organizacionais.

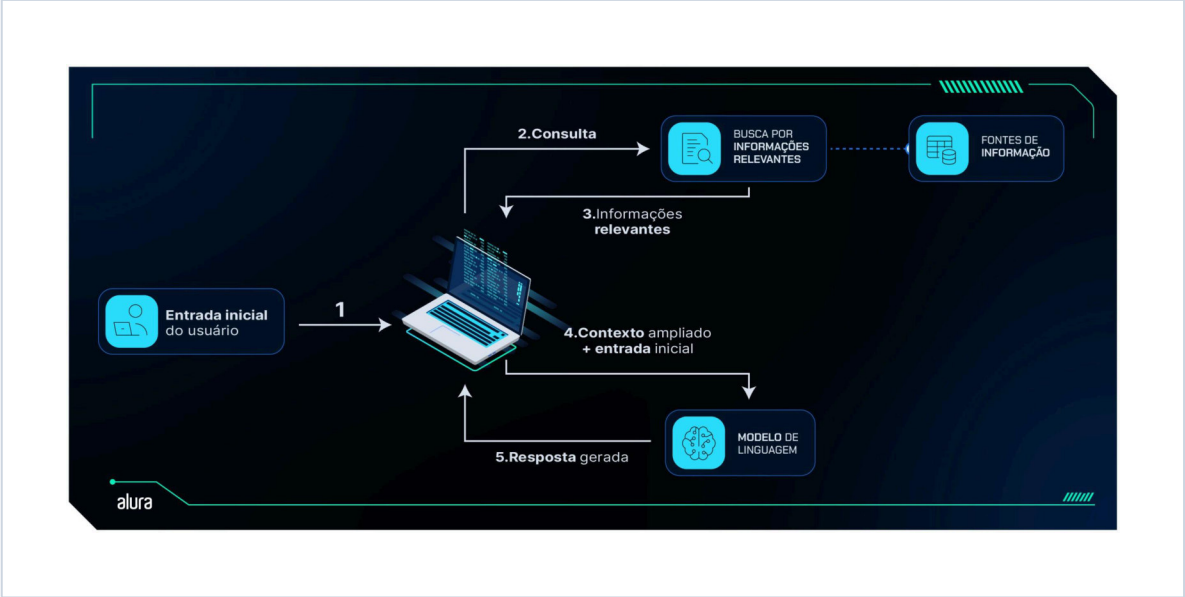
Essas bases são traduzidas para representações matemáticas chamadas *embeddings* (vetores).

Essas representações permitem que o modelo compare a consulta do usuário com os dados disponíveis e identifique as informações mais relevantes com base em cálculos de similaridade vetorial.

Um chatbot de RH, por exemplo, pode buscar diretamente as políticas específicas de férias ao responder a perguntas como “Quantos dias de férias eu tenho?”.

Em seguida, acontece o **enriquecimento do contexto**: as informações recuperadas são integradas ao prompt do usuário para que o modelo gere respostas mais específicas e contextualizadas.

Esses dados externos podem ser atualizados constantemente, seja em tempo real ou por processos periódicos, garantindo que o conhecimento utilizado esteja sempre alinhado ao momento atual.



Principais benefícios do RAG

A adoção do RAG traz uma série de vantagens que a tornam essencial para organizações e desenvolvedores de IA:

- **Redução de custos operacionais**:Treinar novamente um LLM para cada nova atualização de conhecimento é um processo caro e demorado. O RAG elimina essa necessidade ao integrar dados externos, tornando o uso de IA generativa mais acessível.
- **Informações atualizadas**: Conectar o modelo a fontes de dados em tempo real permite que ele forneça respostas baseadas nos acontecimentos mais recentes, algo essencial para áreas como jornalismo, atendimento ao cliente e análises financeiras.
- **Aumento da confiabilidade**: Com a capacidade de citar fontes e oferecer transparência, o sistema ganha a confiança dos usuários. A rastreabilidade das informações também é um ponto chave para aplicações corporativas e científicas.
- **Controle personalizado**: Os desenvolvedores podem ajustar as fontes de conhecimento utilizadas pelo modelo, garantindo que ele esteja alinhado aos objetivos da organização e evitando a exposição de informações sensíveis.

Exemplos de aplicação do RAG

Com essa técnica, é possível ter um sistema com IA que seja seguro e confiável em qualquer contexto.

No ambiente corporativo, por exemplo, **assistentes virtuais** equipados com RAG podem acessar políticas internas, históricos de suporte e bases de conhecimento técnico para gerar respostas precisas para funcionários.

Já na área da educação, **sistemas de apoio a alunos e professores** podem buscar em materiais didáticos atualizados, artigos científicos e guias de estudo, auxiliando a criar planos de aula ou a tirar dúvidas.

Em **comércio eletrônico**, chatbots podem fornecer informações detalhadas para os clientes: sejam sobre produtos, para esclarecer políticas de devolução ou acompanhar o status de pedidos.

Já pensou como aplicar RAG pode beneficiar o seu trabalho? Estamos preparando um curso bem bacana com o passo a passo da aplicação de RAG! Acompanhe nossas novidades.

RAG não é apenas uma solução técnica, mas uma mudança na forma como os modelos de IA interagem com o mundo real, tornando-os mais úteis e confiáveis. Está pronto para explorar as possibilidades?

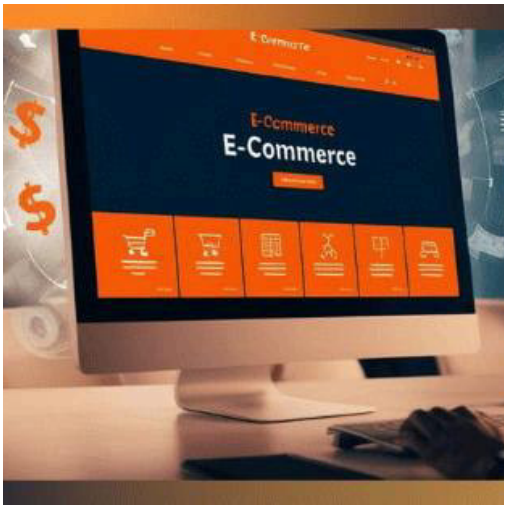
Domine RAG e agentes de IA em menos de 2 horas!

Em poucas horas, você vai **dominar LangChain, LangGraph, Llama 3 e Hugging Face** e aprender a conectar LLMs a dados reais, criar agentes inteligentes e evitar alucinações. Tudo de forma prática, para que você saia com um **projeto prático para seu portfólio**. Além de acelerar seu aprendizado, você ainda garante um **certificado da Alura** para fortalecer seu perfil profissional.

O Flash Skills **Domine RAG e Agentes de IA: Crie agentes, tome decisões e transforme sua carreira** foi feito para quem tem pouco tempo, quer um aprendizado direto ao ponto e está buscando um curso acessível e prático para aplicar no dia-a-dia.

Acesse agora a página do flash skills e matricule-se!

Leia também



Estratégias de SEO para potencializar negócios digitais

As 7 melhores ferramentas de IA que criam imagens

Cor ges

‹ ›

Veja outros artigos sobre Inteligência Artificial

Quer mergulhar em tecnologia e aprendizagem?

Receba conteúdos, dicas, notícias, inovações e tendências sobre o mercado tech diretamente na sua caixa de entrada.

bins.br@gmail.com

ENVIAR

Nossas redes e apps

ENVIAR

bins.br@gmail.com

CURSOS

Cursos de Programação
Lógica | Python | PHP | Java | .NET | Node JS | C | Computação | Jogos | IoT

Cursos de Front-end
HTML, CSS | React | Angular | JavaScript | jQuery

Cursos de Data Science
Ciência de dados | BI | SQL e Banco de Dados | Excel | Machine Learning | NoSQL | Estatística

Cursos de Inteligência Artificial
IA para Programação | IA para Dados

Cursos de DevOps
AWS | Azure | Docker | Segurança | IaC | Linux

Cursos de UX & Design
Usabilidade e UX | Vídeo e Motion | 3D

Cursos de Mobile
Flutter | iOS e Swift | Android, Kotlin | Jogos

Cursos de Inovação & Gestão
Métodos Ágeis | Softskills | Liderança e Gestão | Startups | Vendas

CURSOS UNIVERSITÁRIOS FIAP

Graduação | Pós-graduação | MBA