Usando o Aumento da Geração por Recuperação (RAG) para melhorar a experiência do cliente

ATENTO

Conteúdo

Introdução

Em um mundo onde a inteligência artificial se tornou o novo normal, a **Geração Aumentada por Recuperação (RAG)** surge como uma tecnologia transformadora, projetada para **melhorar a precisão e a confiabilidade de grandes modelos de linguagem (LLMs), incorporando fontes externas de conhecimento.** Na Atento, aproveitamos o RAG para melhorar significativamente nosso Knowledge Assistant, fornecendo aos usuários respostas precisas e contextualmente relevantes. Este material fornece uma exploração detalhada do RAG, suas vantagens, limitações, práticas recomendadas e os mecanismos que implementamos para otimizar a experiência do usuário.

O que é RAG?

A RAG (Geração Aumentada por Recuperação) é uma tecnologia híbrida de IA que combina recuperação de informações com geração de texto para fornecer respostas precisas às consultas do usuário. Esse processo pode ser dividido em duas etapas principais:

PASSO 01

Recuperação de informações

O sistema procura documentos relevantes ou fragmentos de texto de um grande banco de dados em resposta a uma consulta do usuário.

PASSO 02

Geração de texto

A partir dos documentos recuperados, o sistema gera uma resposta informativa adaptada ao contexto da consulta.

Essa abordagem permite uma interação dinâmica, na qual o sistema **aproveita as informações existentes para melhorar a qualidade e a relevância das respostas geradas.**

ATENTO

Benefícios do RAG:

MAIOR PRECISÃO E RELEVÂNCIA

Ao acessar informações atualizadas, o RAG melhora a precisão e a relevância das respostas.

ACESSO ÀS INFORMAÇÕES ATUAIS

O RAG fornece acesso a informações além do limite de treinamento do LLM, garantindo que as respostas sejam baseadas nos dados mais recentes.

CREDIBILIDADE E CONFIANÇA

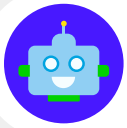
A capacidade de citar fontes melhora a credibilidade e a confiabilidade das informações fornecidas.

CUSTO-BENEFÍCIO

Melhorar os recursos de LLM por meio do RAG pode ser mais econômico do que treinar novamente modelos com novos dados.

ATENTO

Aplicações do RAG:



Chatbots e IA conversacional: Melhora o desempenho do chatbot fornecendo respostas mais abrangentes e contextuais.

Assistência Profissional: Auxilia profissionais com conhecimento especializado.

Suporte ao cliente: Melhora a qualidade do suporte, fornecendo informações precisas e relevantes.

Limitações e práticas recomendadas das soluções RAG

Limitações:

- **Dependência de dados:** A eficácia do RAG depende da qualidade e integridade da base de conhecimento externa.
- **Complexidade da pergunta:** Perguntas complexas ou ambíguas podem testar a capacidade do sistema de recuperar informações precisas.
- **Custo computacional:** O processo duplo de recuperação e geração pode ser computacionalmente intensivo.
- **Desafios da geração de texto:** Se os dados recuperados não forem bem selecionados, podem ocorrer problemas como gerar informações incorretas ou "alucinações".

Práticas recomendadas:

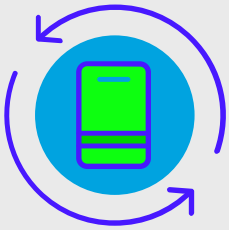
- **Preparação de documentos:** Mantenha um banco de dados atualizado e bem organizado com seções hierárquicas e tags relevantes.
- **Formulação de perguntas:** Formule perguntas de maneira clara e específica para evitar ambiguidade e melhorar a precisão.
- **Melhoria contínua:** Selecione e atualize regularmente o banco de dados para alinhá-lo com as necessidades específicas do sistema RAG.

Mecanismos integrados à solução RAG da Atento

Para melhorar o desempenho e a experiência do usuário do nosso sistema RAG, a Atento integrou vários mecanismos avançados:

Interface amigável

Interfaces que permitem aos usuários controlar vários aspectos da solução com o mínimo de atrito, incluindo alternância dinâmica entre bases de conhecimento em uma única conversa.



Abordagens de reclassificação

Após a recuperação inicial, uma etapa de reclassificação opcional reorganiza fragmentos de texto com base em critérios ou metadados adicionais, melhorando a relevância.

Reformulação de perguntas

O sistema pode criar perguntas semelhantes à consulta do usuário, melhorando a precisão das respostas ao considerar formulações de perguntas diversificadas.

Integração com sistemas existentes

Integração perfeita com outros sistemas e canais de comunicação da empresa, como Microsoft Teams e Hubbie, melhorando a flexibilidade operacional.

Conclusões

A integração da tecnologia de **Geração Aumentada por Recuperação (RAG)** em sistemas de IA oferece uma abordagem transformadora para **melhorar a precisão, relevância e confiabilidade das respostas geradas por grandes modelos de linguagem (LLMs).** Ao alavancar fontes de conhecimento externas, o RAG aborda as principais limitações dos LLMs tradicionais, como informações desatualizadas ou incorretas.

A capacidade do RAG de acessar informações atualizadas garante que as respostas sejam baseadas nos dados mais atualizados disponíveis, melhorando a **experiência geral do usuário.** Além disso, a incorporação de técnicas de reclassificação e reformulação de perguntas refina ainda mais a precisão e a relevância contextual das respostas.

No entanto, a implementação bem-sucedida do RAG requer preparação meticulosa de documentos, formulação clara de perguntas e curadoria contínua do banco de dados para manter a eficácia do sistema. Compreender essas práticas recomendadas é essencial para maximizar o potencial do RAG e garantir sua escalabilidade e eficiência.

Vamos começar?

Entre em contato conosco hoje para saber como nossas soluções avançadas RAG podem melhorar suas operações comerciais e fornecer aos seus clientes um serviço incomparável.

Atento.com