alura

LangChain: Criando chatbots inteligentes com RAG

Cadeias de conversação robusta

MERGULHE EM TECNOLOGIA_

Instrutor(a): Leonardo Pena

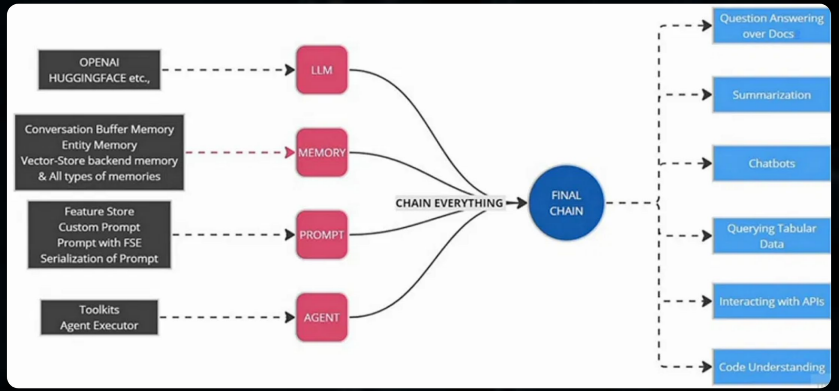
Seja bem vindo (a)!

Aula

Cadeias de Conversação Robusta

O que é ConversationalRetrievalChain?

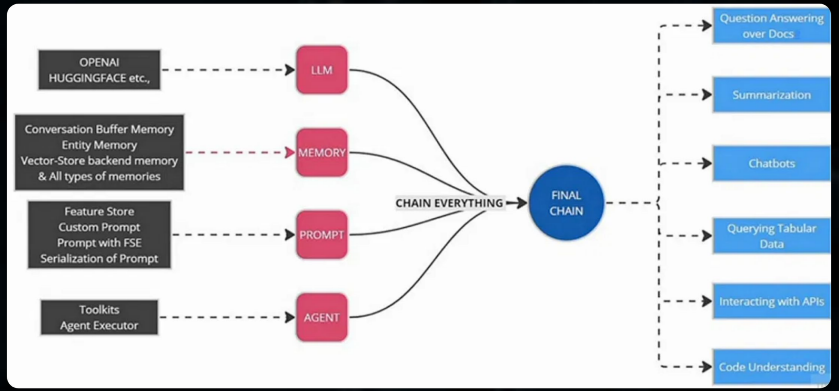
ConversationalRetrievalChain é um tipo de cadeia avançada no LangChain que gerencia o histórico da conversa para manter o contexto em um diálogo interativo.



O que é ConversationalRetrievalChain?

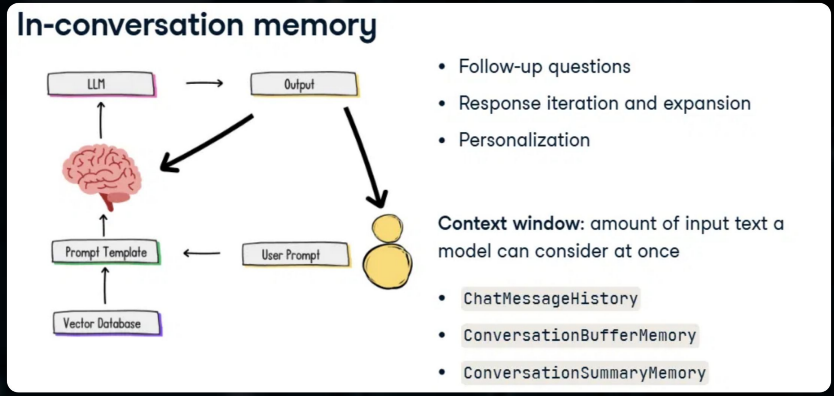
- Utiliza o histórico de conversas anteriores para contextualizar novas perguntas
- Reformula a pergunta atual em uma nova pergunta autônoma e contextualizada
- Realiza busca nos documentos com a pergunta reformulada para maior precisão

Esta abordagem permite que o chatbot **mantenha o contexto** ao longo de múltiplas interações, criando uma experiência de conversa mais natural e Inteligente.



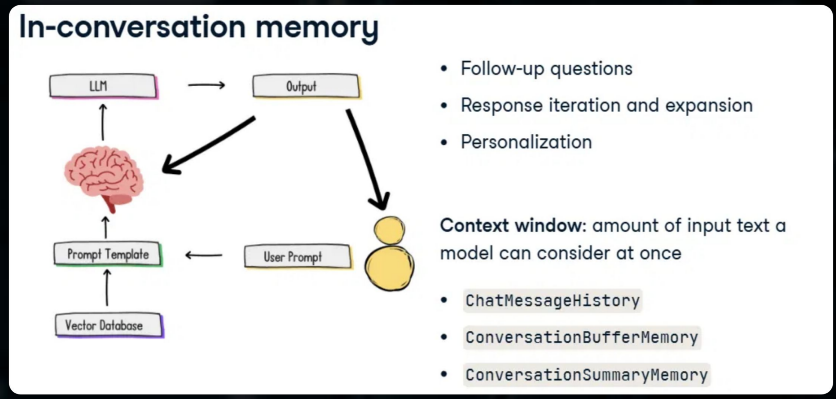
Gerenciamento de Memória

O **gerenciamento de memória** é um aspecto crítico para chatbots conversacionais que permite "lembrar" interações passadas e manter o contexto ao longo do tempo.



Gerenciamento de Memória

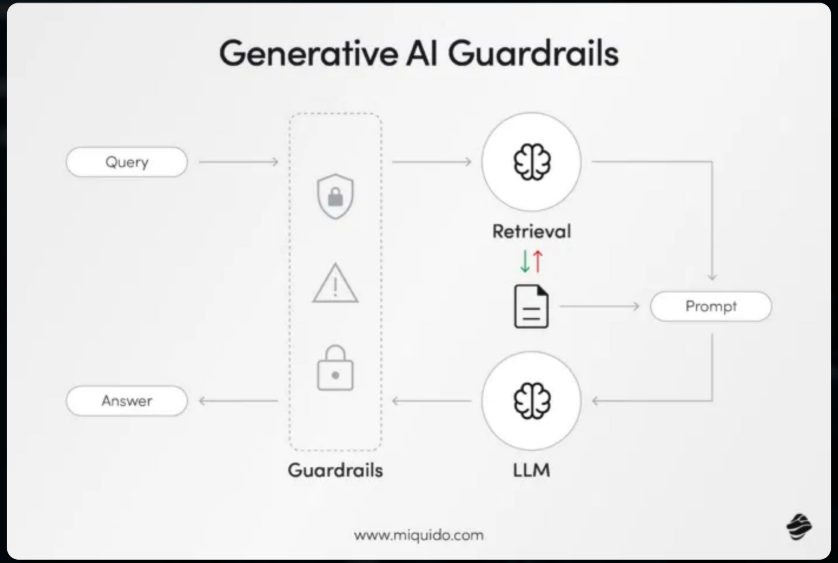
- **ConversationBufferMemory**
  Armazena todas as interações anteriores, mas pode consumir muitos tokens em conversas longas.
- **ConversationBufferWindowMemory**
  Armazena apenas as últimas k interações, equilibrando contexto e consumo de tokens.
- **ConversationSummaryMemory**
  Mantém um resumo das interações anteriores em vez do histórico completo.



A escolha do tipo de memória deve considerar o equilíbrio entre contexto e eficiência, adaptando-se ao caso de uso específico do chatbot.

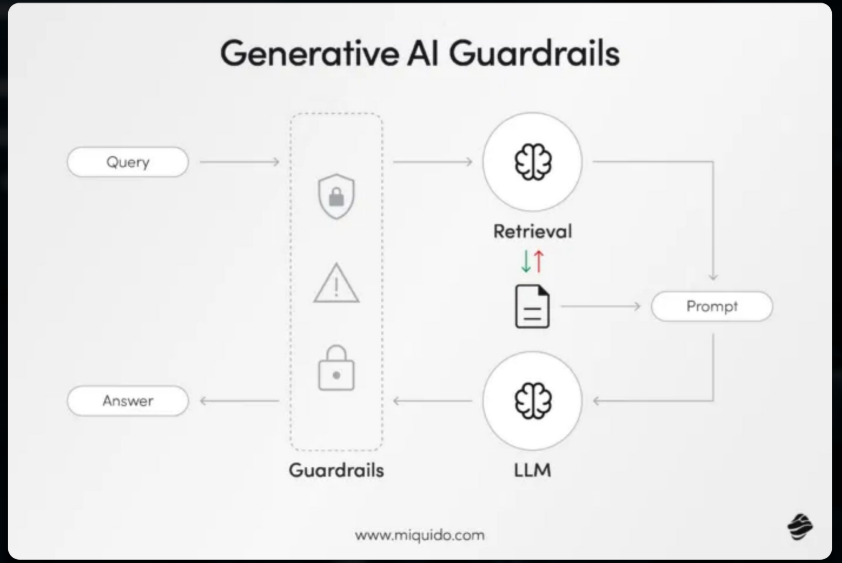
Guardrails de Segurança

Guardrails de segurança são mecanismos implementados como filtros para garantir que o chatbot opere dentro de limites definidos, prevenindo a geração de respostas indesejadas.



Guardrails de Segurança

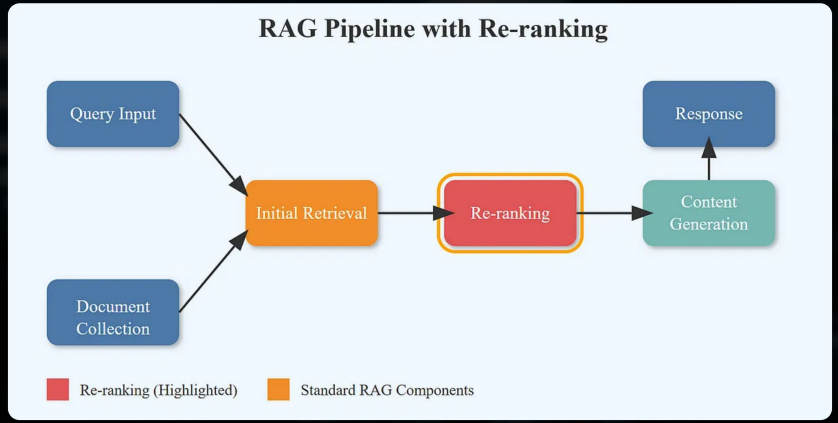
- **Filtros de conteúdo tóxico**
  Previnem respostas ofensivas, discriminatórias ou inadequadas.
- **Proteção de informações pessoais (PII)**
  Evitam a exposição de dados sensíveis como CPF, endereços ou informações financeiras.
- **Limitadores de escopo**
  Garantem que as respostas permaneçam dentro do domínio específico da aplicação.



Implementar guardrails é essencial para aplicações em produção, protegendo usuários, dados e a reputação da empresa contra riscos associados a LLMS.

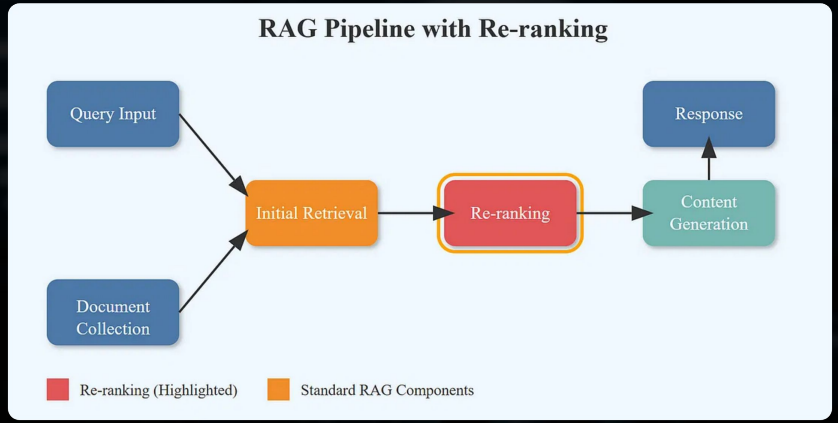
Re-ranking

Re-ranking é uma técnica que refina os resultados da etapa de recuperação inicial, reordenando os documentos recuperados com base em uma avaliação mais sofisticada de sua relevância contextual.



Re-ranking

1. **Recuperação Inicial**
   Documentos são recuperados usando métodos tradicionais de similaridade vetorial.
2. **Avaliação Contextual**
   Um modelo de re-ranking (como Cohere Rerank) analisa a relevância semântica profunda entre a consulta e cada documento.
3. **Reordenação**
   Os documentos são reordenados com base nesta avaliação mais precisa



Benefícios: maior precisão nas respostas, redução de alucinações do LLM e melhor experiência para o usuário final.