Como o uso de RAG melhora a contextualização?

Técnicas Rag para o melhorar o Few-shot

O que é RAG?

Retrieval-Augmented Generation

01

Busca por similaridade

02

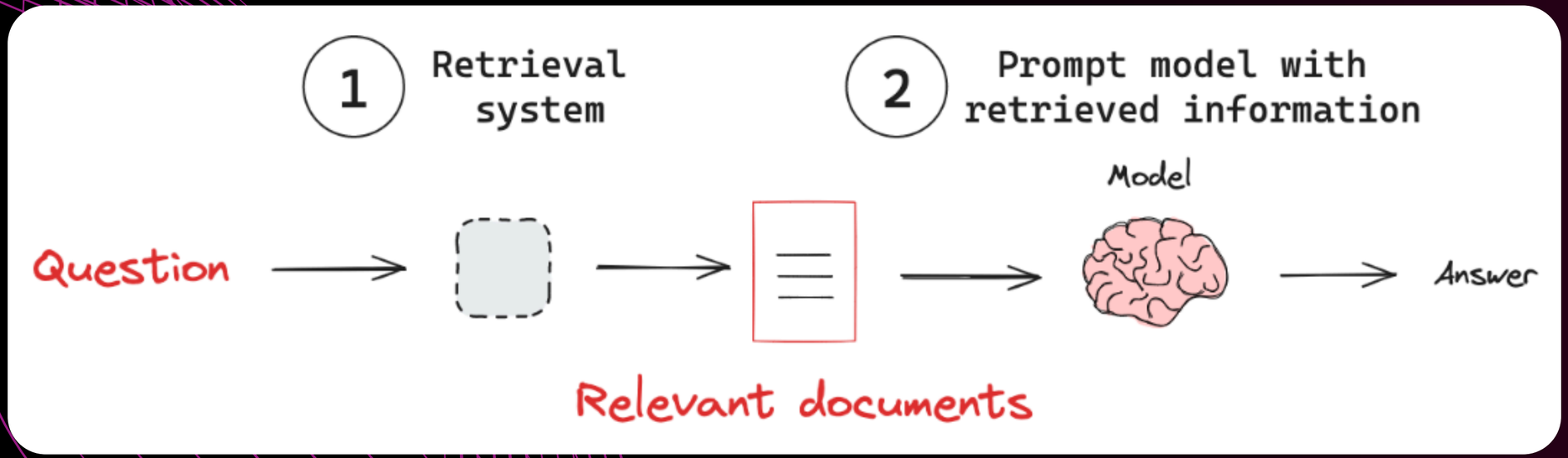
Maior precisão contextual

03

Uso de bases de conhecimento externas

04

Escalabilidade

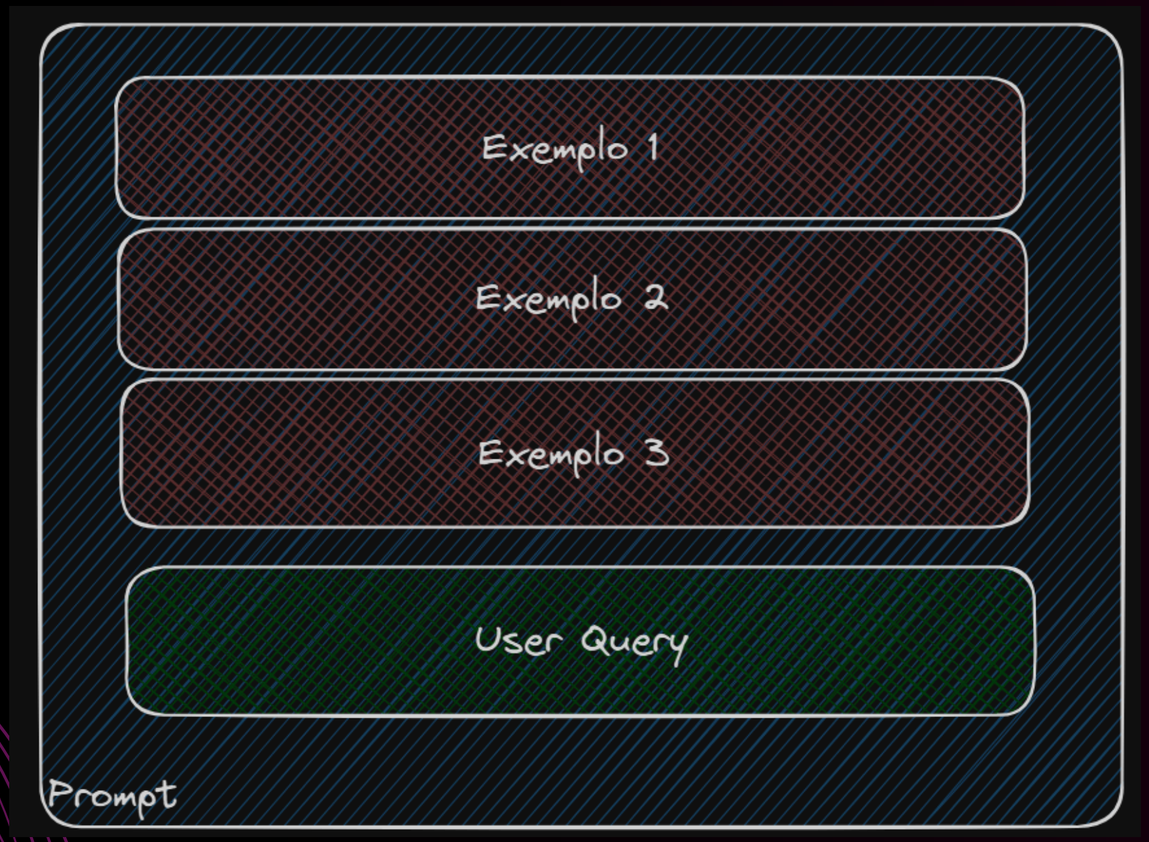


Limitações do Few-shot

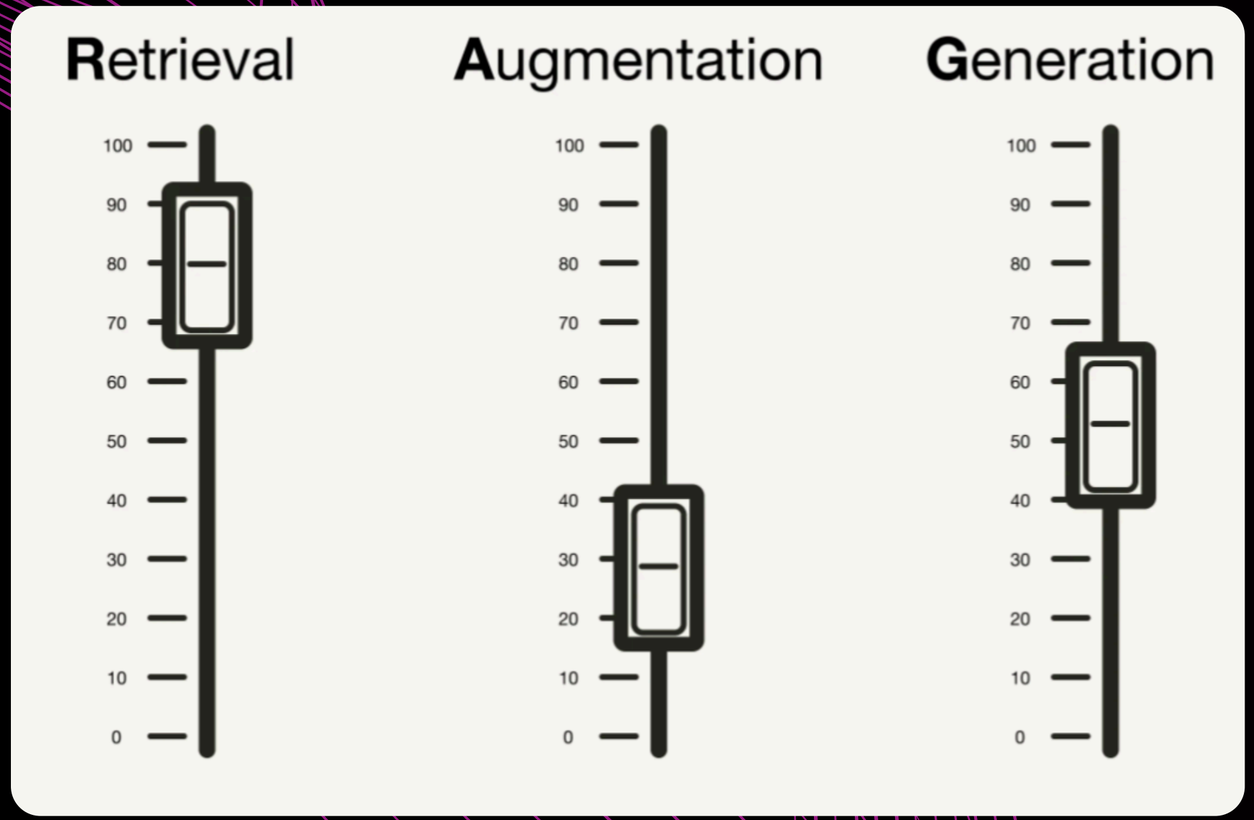
Eficiência para Domínios Específicos

Falta de contexto detalhado

Exemplos Estáticos



RAG como solução para o Few-shot



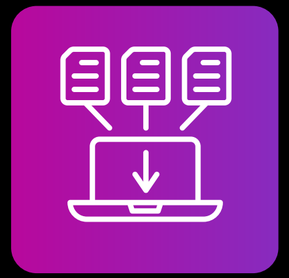
01 **Recuperação de conhecimento relevante**

02 **Domínio baseado nos documentos de referência**

03 **Exemplos Dinâmincos**

Como o RAG funciona?

RAG combina busca de informações e geração de linguagem natural para criar respostas precisas usando dados relevantes e contexto específico.



Retrieval

Localiza informações relevantes em uma base de conhecimento para responder à consulta.

Augmentation

Enriquece o modelo com os dados recuperados, adicionando contexto específico.

Generation

Gera uma resposta final em linguagem natural combinando conhecimento pré-treinado e contexto recuperado.

RAG-Fusion

Uma técnica avançada de RAG

RAG Fusion integra múltiplas fontes de dados recuperados antes da geração, combinando informações para criar respostas mais completas e precisas.

Combinação de Fontes

Recupera dados de várias fontes relevantes simultaneamente

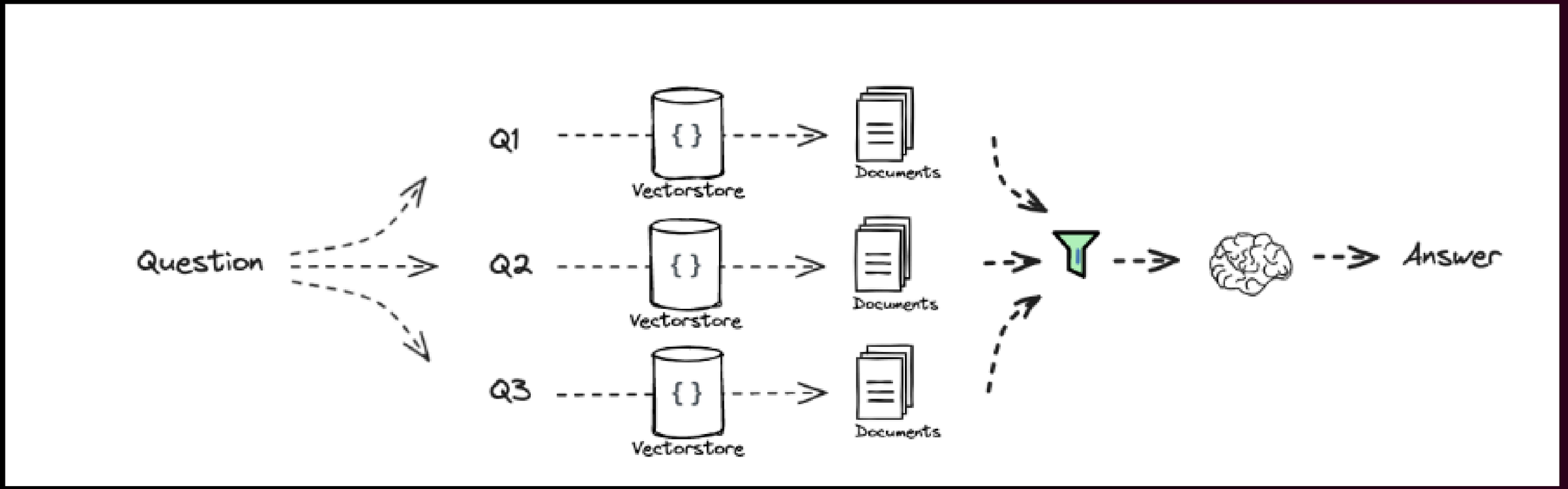
Integração de Contexto

Une as informações recuperadas para enriquecer o contexto antes da geração

Resposta Aprimorada

Gera uma resposta mais completa ao combinar insights de múltiplas bases de dados

Explicação prática do funcionamento de RAG Fusion



Explicação
Técnica + Exemplos