alura

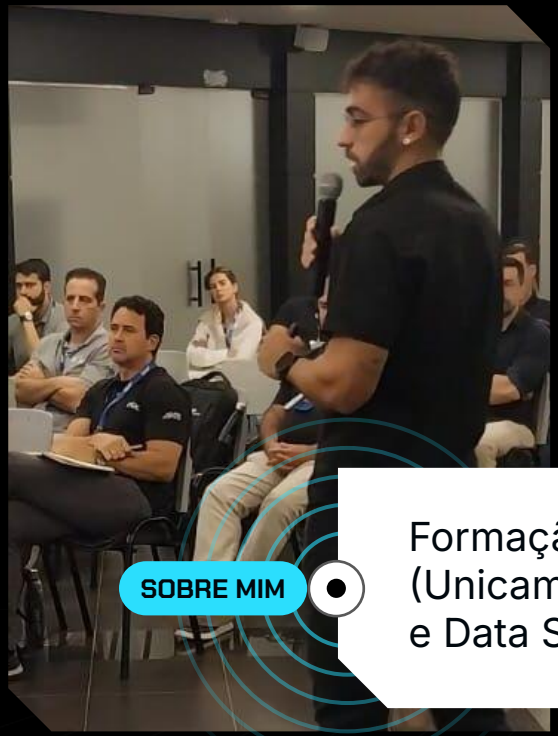
LangChain: Criando chatbots inteligentes com RAG

Arquitetura RAG na Prática

MERGULHE EM TECNOLOGIA_

Instrutor(a): Leonardo Pena

Seja bem vindo (a)!



SOBRE MIM

Leonardo Pena

Cientista de Dados Sr e Lead

Formação em Estatística (Unicamp) com pós em IA e Data Science

+12 anos de vivência no mundo de dados. Hoje, professor e c

alura

VISÃO GERAL DO CURSO

alura

AULA 1

Arquitetura RAG na Prática

AULA 2

Armazenamento Vetorial com FAISS/Chroma

AULA 3

Embeddings de Alta Performance

Pipelines para Dados Complexos

AULA 4

AULA 5

Cadeias de Conversação Robusta

AULA 6

Avaliação com LangSmith & RAGAS

Hybrid Search & Técnicas Avançadas

AULA 7

AULA 8

Deploy na Nuvem

AULA 9

Projeto Capstone

Aula

Arquitetura RAG na Prática

OBJETIVO DA AULA

Vamos entender como fazer a IA responder com base em informações específicas e atualizadas

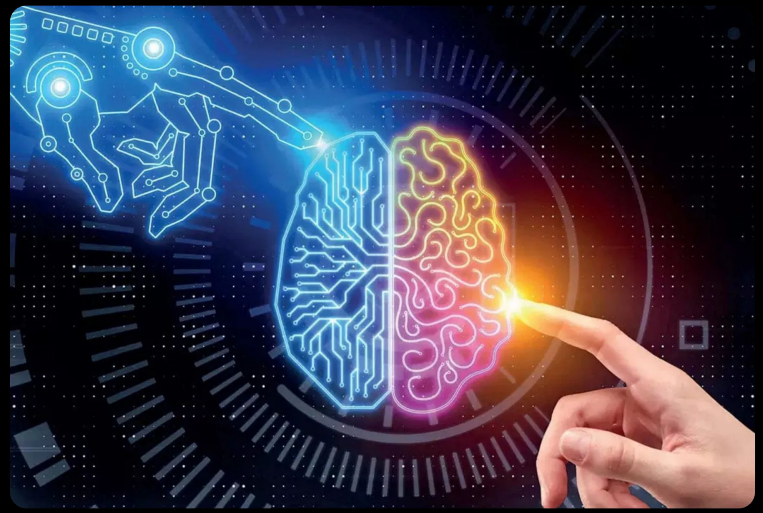
O PROBLEMA: "ALUCINAÇÕES" E CONHECIMENTO DESATUALIZADO

// Modelos de linguagem como GPT são treinados com uma base de dados gigantes, mas **fixa e limitada**. Ou seja...

O PROBLEMA: "ALUCINAÇÕES" E CONHECIMENTO DESATUALIZADO

Conhecimento Desatualizado: Não conhecem eventos recentes ou informações posteriores a data do treinamento

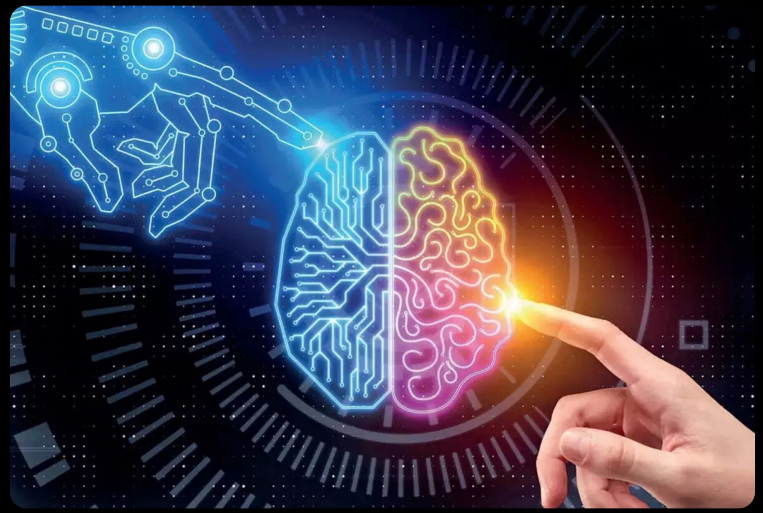
Dados Privados Inacessíveis: Não tem acesso a informações privadas da sua empresa, documentos internos ou bases específicas



O PROBLEMA: "ALUCINAÇÕES" E CONHECIMENTO DESATUALIZADO

Alucinações: Podem "inventar" respostas quando não sabem algo, criando informações que parecem plausíveis mas são falsas

Contexto Limitado: Não conseguem acessar diretamente seus documentos, catálogos de produtos ou bases específicas



A SOLUÇÃO: RAG (RETRIEVAL-AUGMENTED GENERATION)

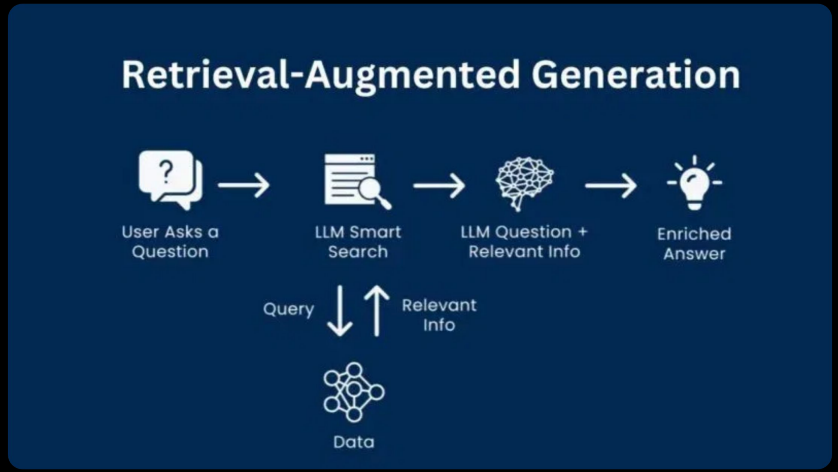
RAG combina o melhor dos dois mundos, a **busca precisa** de um motor de busca com a **capacidade de conversação** de um LLM

Busca (Retrieve) + Geração (Generate) = Respostas Inteligentes e Confiáveis

A SOLUÇÃO: RAG (RETRIEVAL-AUGMENTED GENERATION)

A primeira etapa, "Recuperação", consiste em encontrar os trechos de informação mais relevantes em seus documentos para responder a uma pergunta.

1. O sistema busca os documentos relevantes em sua base de conhecimento
2. Em seguida, aumenta o prompt do usuário, inserindo a informação encontrada como contexto
3. Finalmente, o LLM gera uma resposta com a instrução clara de usar APENAS aquele contexto



COMO FUNCIONA: AUMENTANDO O CONTEXTO

O "truque" do RAG é dar ao LLM um "livro de consulta" específico para cada pergunta, permitindo respostas precisas e baseadas em fatos

Pergunta do Usuário

O usuário faz uma pergunta

Busca Semântica

Sistema busca documentos relevantes

Aumento de Contexto

Documentos adicionados ao prompt

Geração da Resposta

LLM responde baseado no contexto

COMO FUNCIONA: AUMENTANDO O CONTEXTO

Pergunta do Usuário:

"Qual é a política de devolução para produtos eletrônicos?"

Contexto Recuperado:

"Produtos eletrônicos podem ser devolvidos em até 30 dias com a nota fiscal. Itens danificados não são elegíveis."

Resposta Gerada:

"Nossa política permite a devolução de produtos eletrônicos em até 30 dias, desde que você apresente a nota fiscal e o produto não esteja danificado."

COMPONENTES ESSENCIAIS DO RAG

Embeddings

Representações numéricas (vetores) que capturam o significado semântico dos textos, permitindo busca por similaridade. São o "DNA" da informação.

Banco de Dados Vetorial

Armazena e indexa os embeddings para busca rápida e eficiente. Exemplos: FAISS (Facebook), Chroma, Pinecone, Weaviate.

Chunking

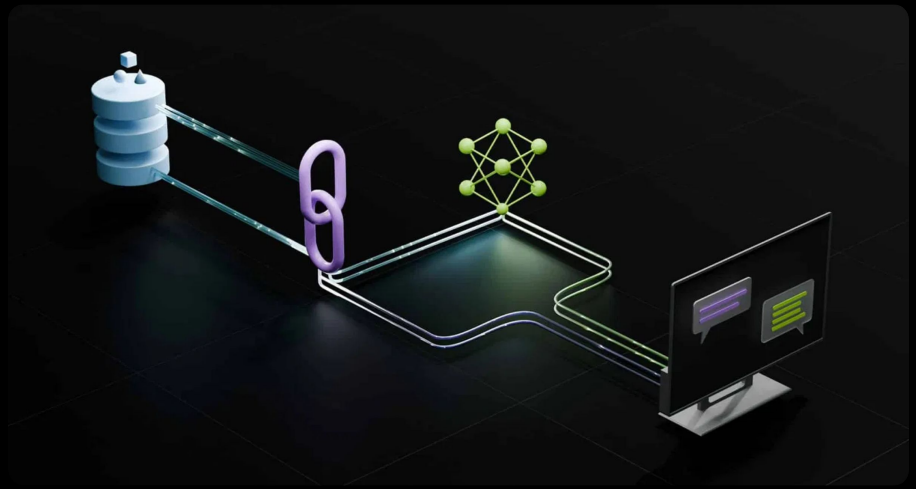
Processo de dividir documentos em pedaços menores e coerentes para processamento e recuperação eficientes. Crucial para a qualidade das respostas.

Modelo de Linguagem (LLM)

O "cérebro" que gera respostas coerentes baseadas no contexto recuperado. Pode ser local ou via API (OpenAI, Anthropic, etc).

FLUXO COMPLETO DO RAG

O ciclo de vida completo de um sistema RAG envolve duas fases principais: **Ingestão de Dados** e **Consulta**



1. **Ingestão de Documentos**

   Documentos são coletados, processados e divididos em chunks menores para facilitar a recuperação.

2. **Geração de Embeddings**

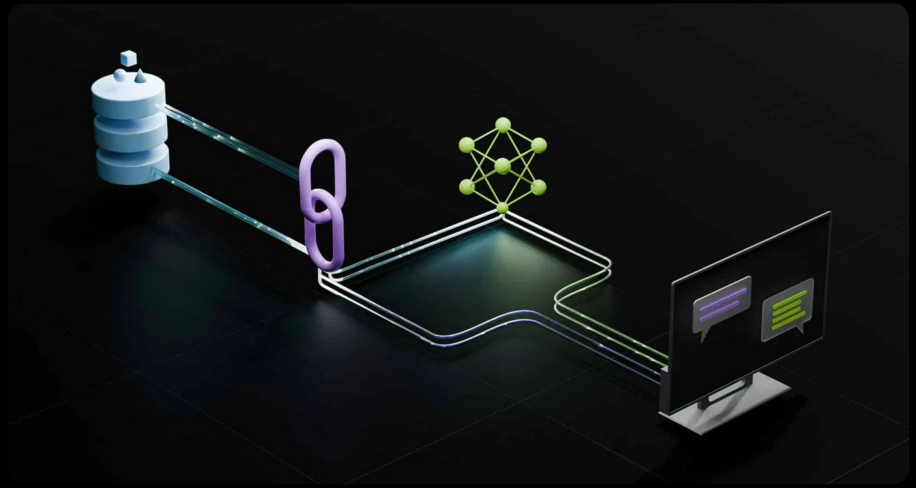
   Cada chunk é convertido em um vetor numérico (embedding) que representa seu significado semântico.

3. **Indexação Vetorial**

   Os embeddings são armazenados em um banco de dados vetorial para busca eficiente.

FLUXO COMPLETO DO RAG

O ciclo de vida completo de um sistema RAG envolve duas fases principais: **Ingestão de Dados** e **Consulta**



4 **Consulta e Recuperação**

A pergunta do usuário é convertida em embedding e usada para buscar chunks relevantes.

5 **Geração da Resposta**

O LLM gera uma resposta baseada na pergunta e nos chunks recuperados como contexto.

POR QUE RAG É A REVOLUÇÃO

Reduz Alucinações

As respostas são baseadas em fontes concretas e verificáveis, aumentando a confiabilidade.

Conhecimento Atualizado

Para atualizar a IA, basta atualizar os documentos na base de conhecimento.

Custo-Benefício

Muito mais barato e rápido do que treinar um modelo do zero (fine-tuning) para cada nova informação.

Transparência

É possível saber exatamente qual fonte foi usada para gerar uma resposta, facilitando a auditoria.

Privacidade

Seus dados sensíveis podem permanecer dentro da sua infraestrutura, sem exposição externa.

CASOS DE USO DO RAG

Atendimento ao Cliente

Chatbots e assistentes virtuais que respondem com base em manuais, FAQs e histórico de atendimentos.

> *"Como posso trocar a bateria do meu dispositivo modelo X200?"*

Suporte Técnico e Médico

Assistentes que ajudam profissionais a encontrar informações em grandes bases de conhecimento técnico.

> *"Quais são os efeitos colaterais conhecidos da interação entre medicamentos A e B?"*

Análise de Documentos Legais

Extração de informações específicas de contratos, termos e documentos jurídicos complexos.

> *"Quais são as cláusulas de rescisão antecipada nos contratos da empresa X?"*

Pesquisa e Desenvolvimento

Acesso rápido a informações em artigos científicos, patentes e documentação técnica interna.

> *"Quais pesquisas recentes abordam o uso de grafeno em baterias de lítio?"*