Fundamentando um copiloto com seus próprios dados

Otimizar Através do Retrieval Augmented Generation (RAG)

Fundamentando um copiloto com seus próprios dados

Os modelos de linguagem criam respostas coerentes às perguntas, mas em que se baseiam essas respostas?

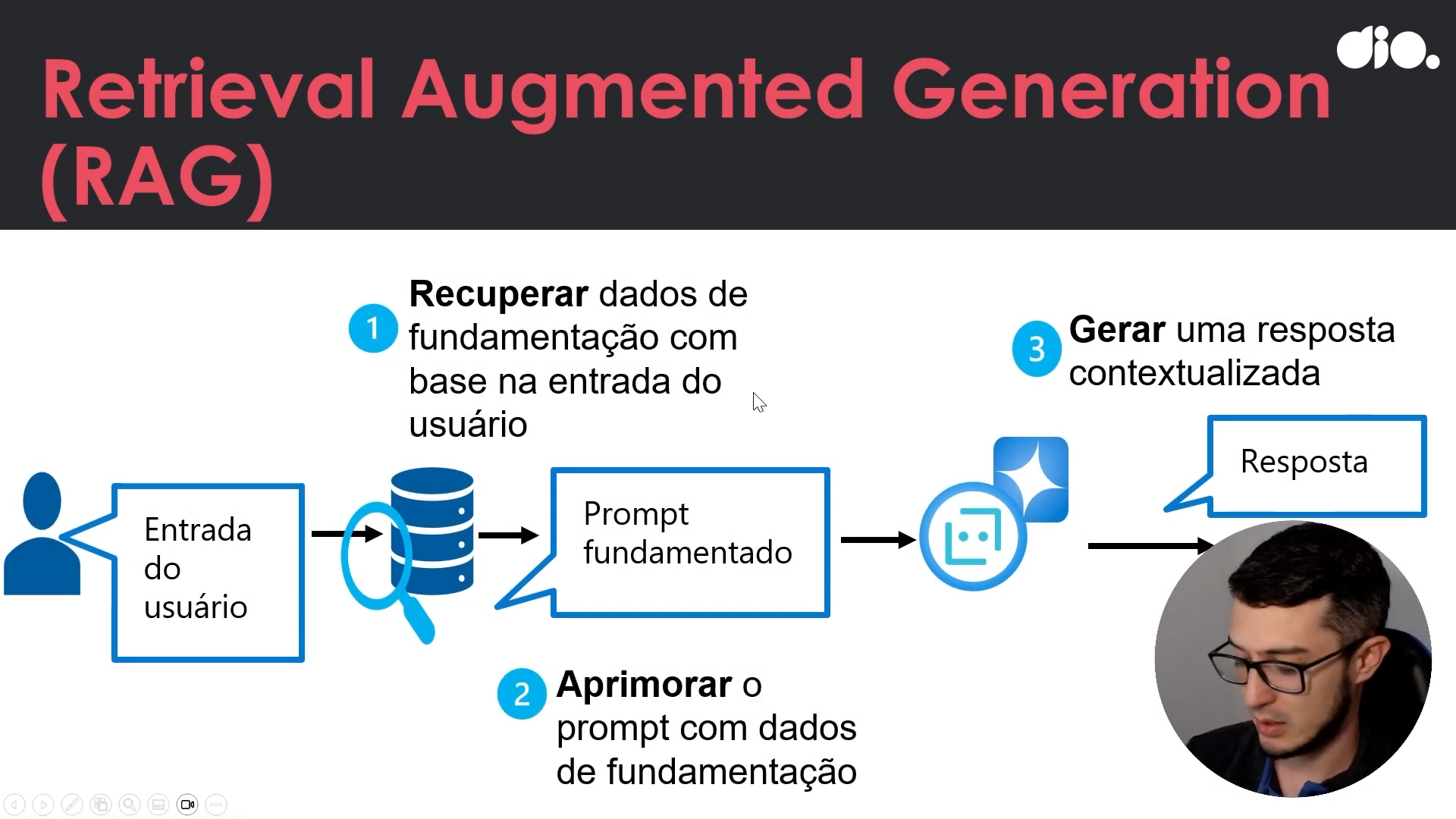
A fundamentação fornece contexto específico ao modelo para fornecer respostas precisas e relevantes.



Retrieval Augmented Genaration (RAG)

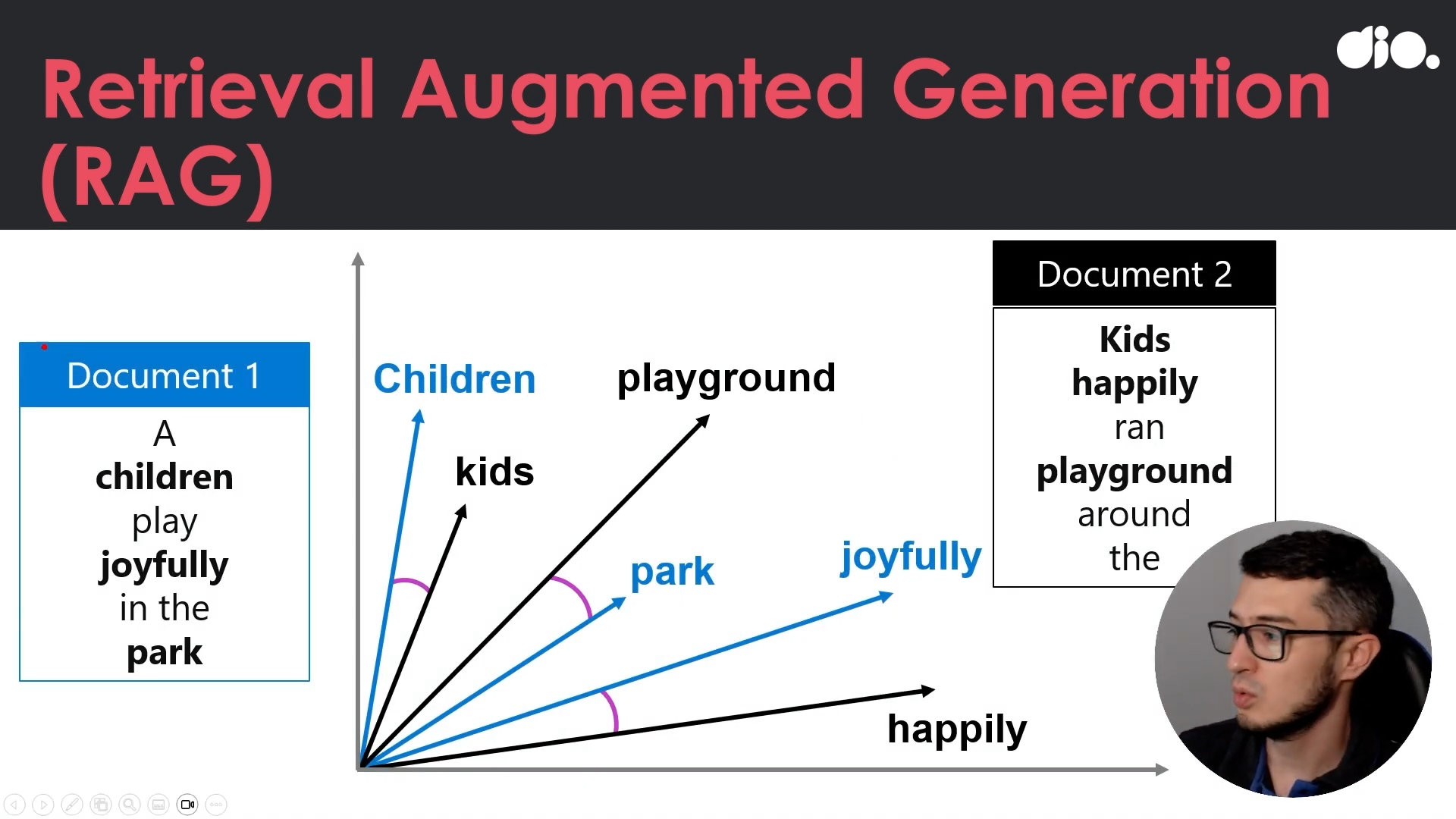
Uma RAG reúne dados relevantes para incluir no prompt para que o modelo de linguagem utilize como contexto de fundamentação.

Um padrão RAG é um design arquitetural para RAG, incluindo dados relevantes recuperados no prompt:

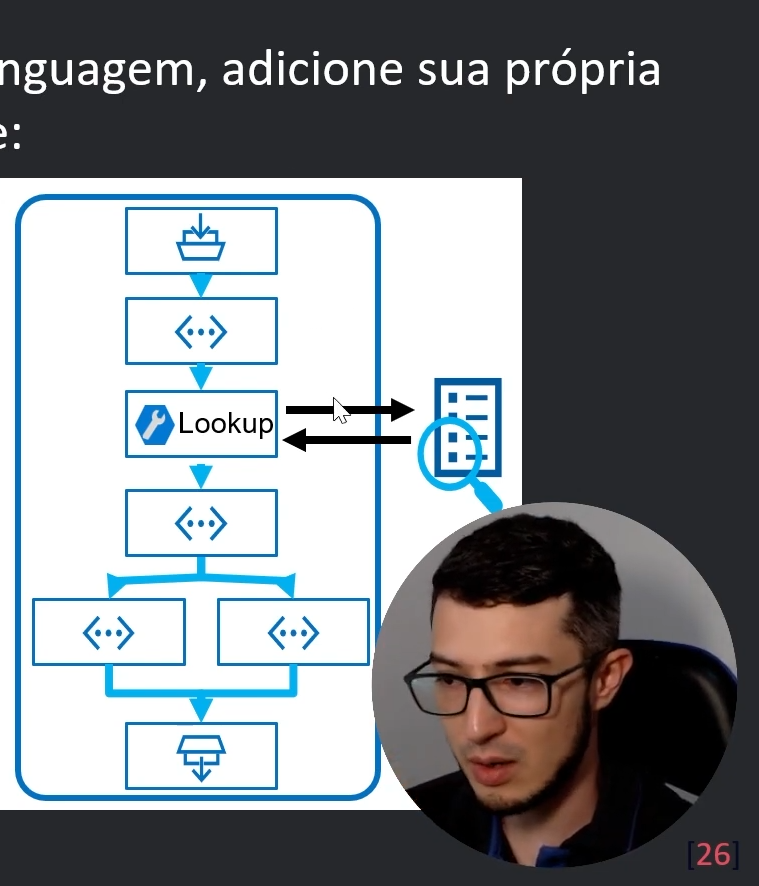


Os dados são armazenados como embeddings em um índice de busca baseado em vetores.

Embeddings permitem uma melhor recuperação de dados para similaridades semânticas nos resultados da busca.



Adicione seus dados em fluxo de pompt

Para fundamentar as respostas do seu modelo de linguagem, adicione sua própria fonte de dados para recuperar o contexto relevante:

1. Adicione seus dados a um projeto Azure AI.
2. Indexe seus dados com o Azure AI Search .
3. Consulte seus dados indexados em um fluxo de prompt com a ferramenta Idenx Lookup.
4. Referencie o contexto recuperado no prompt.
5. Envie o prompt com o contexto para um LLM.

A resposta gerada será fundamentada pelo contexto recuperado.

Materiais de Apoio

Os materiais complementares e de apoio que oferecemos têm como objetivo fornecer informações para facilitar e enriquecer a sua jornada de aprendizado no curso "Otimizar Através do Retrieval Augmented Generation (RAG)". Aqui você encontrará links úteis, como slides, repositórios e páginas oficiais, além de dicas sobre como se destacar na DIO e no mercado de trabalho 😉

Recursos Adicionais

Durante este conteúdo, compreendemos os fundamentos da engenharia de prompts. Para ajudá-lo a aprofundar o conhecimento, disponibilizamos a seguir o material complementar contendo os conteúdos e links apresentados no curso:

Slide: Módulo 7 - Otimizando Modelos de Linguagem para Aplicações Generativas de IA.pptx

Dicas e Links Úteis

Para se desenvolver ainda mais e se destacar na DIO e no mercado de trabalho, sugerimos os seguintes recursos:

Artigos e Fórum da DIO: Compartilhe seus conhecimentos e dúvidas através dos artigos (visíveis globalmente na plataforma da DIO) e nos fóruns específicos para cada experiência educacional, como nossos Bootcamps.

Rooms: Participe do Rooms, uma ferramenta de bate-papo em tempo real onde você pode interagir com outros participantes dos nossos Bootcamps, compartilhando dúvidas, dicas e snippets de código.

Exploração na Web: Utilize motores de busca para aprofundar seu conhecimento sobre temas específicos. Páginas como o StackOverflow são recursos valiosos para encontrar soluções e expandir seu entendimento.

Com esses materiais complementares, você estará bem equipado para explorar todo o potencial e se destacar em suas iniciativas. Continue aproveitando as oportunidades de aprendizado, e não hesite em buscar mais conhecimento e compartilhar suas descobertas com a comunidade!

