Adicionando contexto aos modelos do Amazon BedRock

Gabriel Bella Martini (ele/dele)

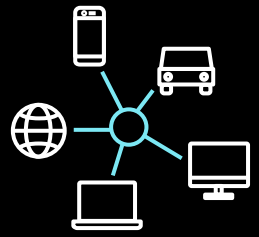
Arquiteto de Soluções para IA/ML

AWS Setor Público

Agenda

- Quais os desafios ao utilizar um LLM (Large Language Model)
- Introdução ao Retrieval Augmented Generation (RAG)
- Knowledge bases para Amazon BedRock
- Demonstração
- Próximos passos

Quais os desafios ao utilizar um LLM

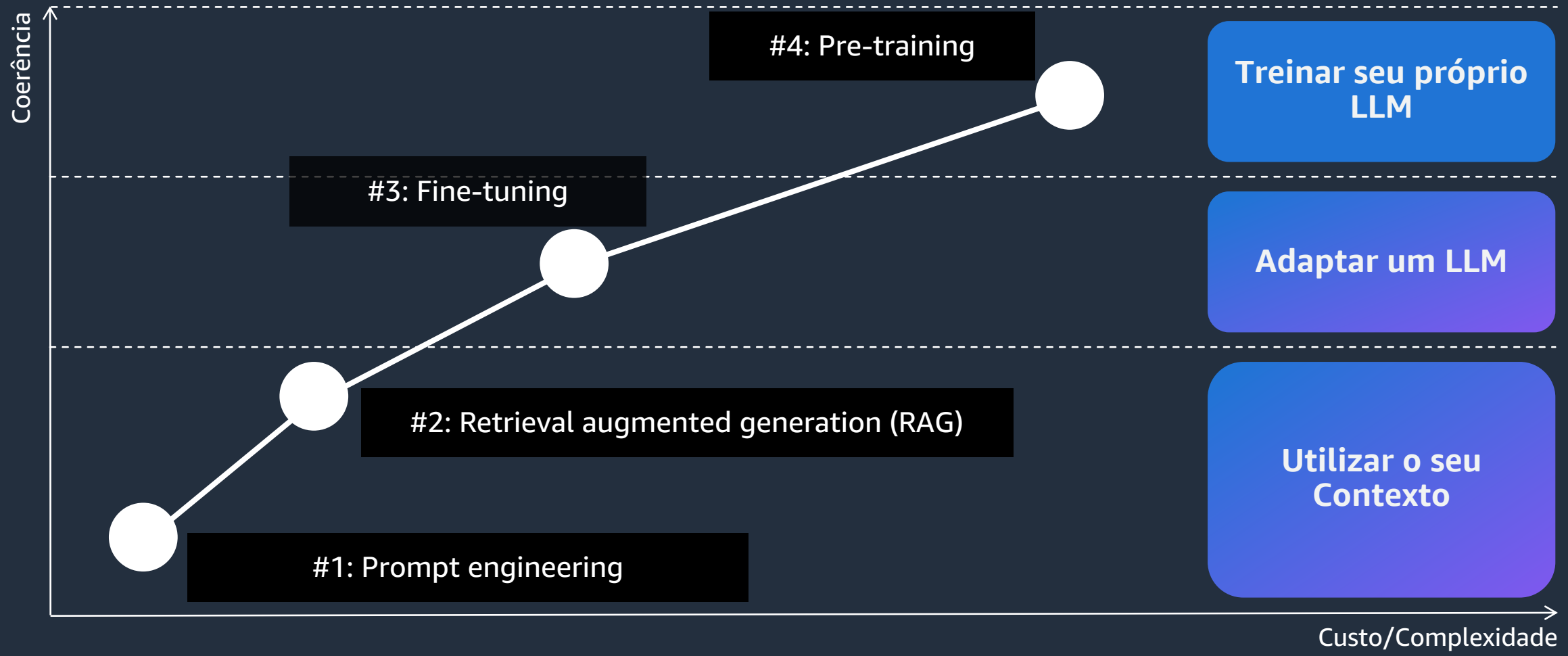


Os dados de treinamento são incompletos

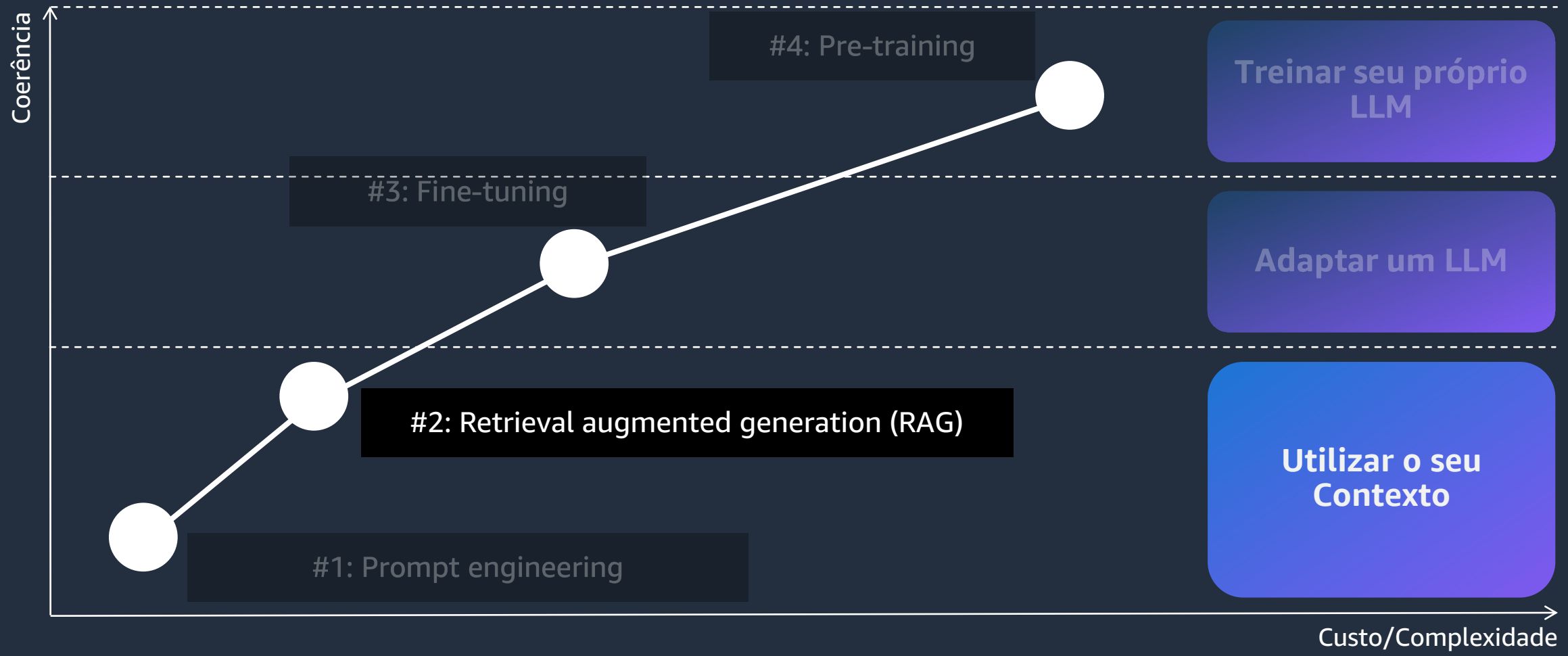
Podem ocorrer alucinações

Como posso ter maior confiança e assertividade

Como customizar



Como customizar



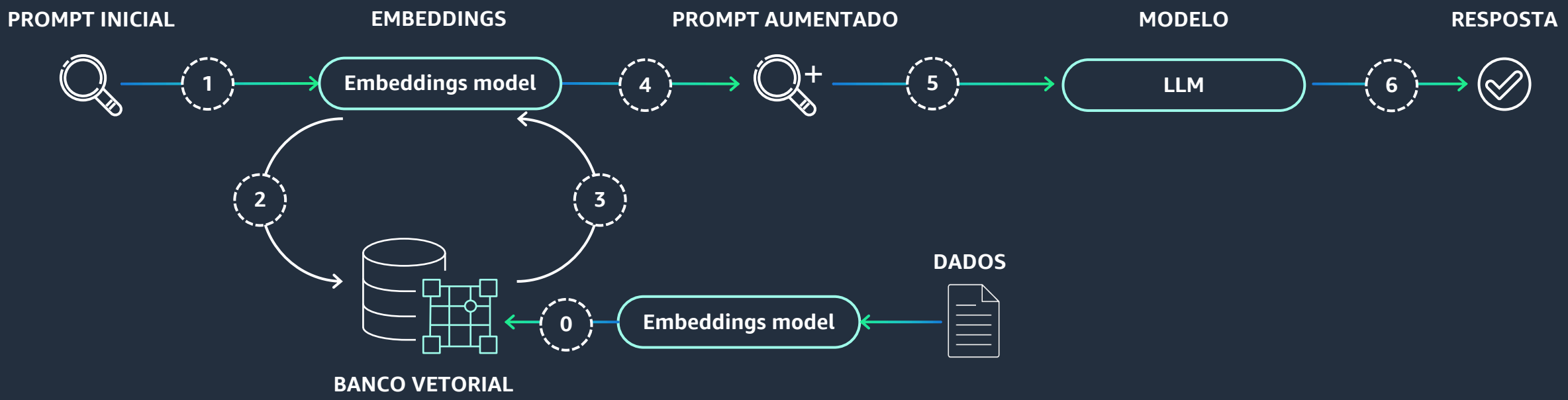
Por que Retrieval Augmented Generation (RAG) é um padrão comum?

Entrega contexto para o LLM além do dado de treinamento

Reduz possíveis alucinações

Aumenta a confiança e traz a fonte da resposta

Não exige conhecimento mais profundo de Machine Learning e operação de modelos



Knowledge bases para Amazon Bedrock

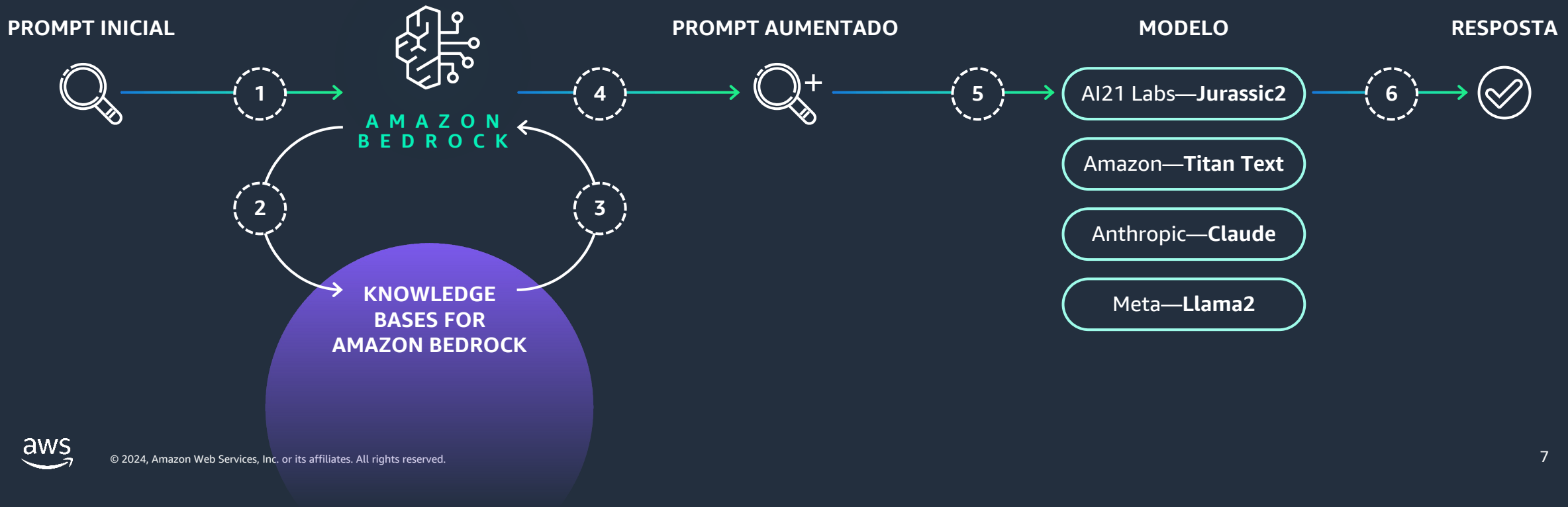
SUPORTE NATIVO PARA RETRIEVAL AUGMENTED GENERATION (RAG)

Conectar de forma segura os LLMs às fontes de dados

Fluxo de trabalho RAG gerenciado, incluindo ingestão, recuperação e aumento

Gerenciamento de contexto de sessão integrado para conversas de várias etapas

Citações automáticas com recuperações para melhorar a transparência



Titan Embeddings

Amazon Titan Embeddings

V2.0

Traduz entradas de texto (palavras, frases) em representações numéricas (embeddings). Similaridade por embeddings produz respostas mais relevantes e contextuais do que a similaridade de palavras.

Máximo de tokens: **8,000**
Vetor de saída**: 1,536**
Línguagem: **Multilíngue** (25 línguas)

ID: *amazon.titan-embed-text-v1*

Destaques

- Titan Embeddings oferece embeddings rápidos, econômicos, de alto desempenho e precisos em 25 idiomas
- Otimizado para tarefas de recuperação de texto, similaridade semântica e agrupamento
- Aplicações deste modelo incluem busca semântica e personalização

ANTHROP\C

Anthropic Claude

Claude

Anthropic oferece a família Claude de grandes modelos de linguagem construídos especialmente para conversas, sumarização, perguntas e respostas, automação de fluxo de trabalho, codificação e muito mais. Claude também pode receber direções sobre personalidade, tom e comportamento.

Máximo de tokens: **200000**
Linguagem: **Multilíngue**
IDs:
- *anthropic.claude-instant-v1*
- *anthropic.claude-v2:1*

Destaques

- Janela de contexto longa (200k) permite processar grandes quantidades de texto de uma vez
- Disponível em dois tamanhos para ajudar a escolher o modelo do tamanho adequado com base em considerações de latência, precisão e custo
- IA Constitucional com foco em segurança

Demonstração

Próximos Passos

Comece a usar o Amazon Bedrock

Faça com um tutorial passo a passo

Mergulhe fundo com um workshop prático

Obrigado!

Gabriel Bella Martini

Arquiteto de Soluções para IA/ML/IA Generativa

in/gabrielbmartini