Modelos de IA Mais Ágeis e Econômicos: Como Especializar LLMs para Utilities

Silvio Nunes
Senior Solutions Architect
aws-power-utilities-brazil@amazon.com

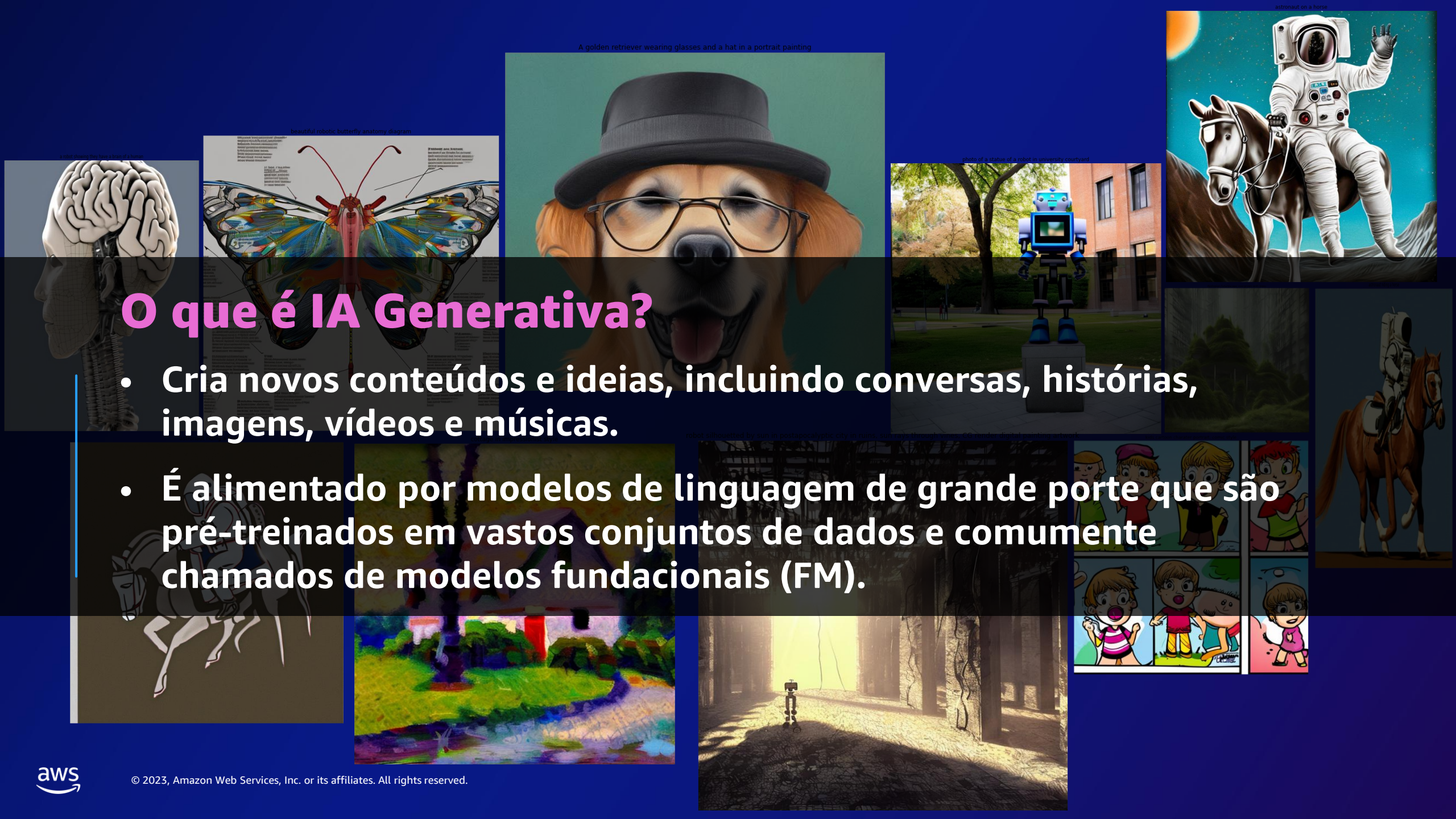
Utilities Telecom & Technology Council América Latina™

Agenda

1. Fundamentos e Desafios
2. Abordagens de customização de modelos
3. Distilação de Modelos: O que é e como funciona
4. Execução de IA distribuída na borda
5. Principais aprendizados e próximos passos

Desafios e Oportunidades para IA Generativa em Utilities

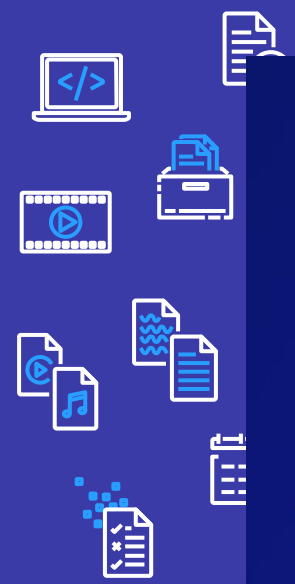
- Empresas do setor de **utilities** estão acelerando o uso de **IA Generativa** para diferentes casos de uso
- Em setores regulados, como utilities, pode ser necessário **customizar os modelos de IA**
- Modelos **LLMs tradicionais** podem ser **caros e lentos** para produção — especialmente com grandes volumes de uso.
- Como equilibrar o uso de IA Generativa com **custo, latência e precisão**?



O que é IA Generativa?

- Cria novos conteúdos e ideias, incluindo conversas, histórias, imagens, vídeos e músicas.
- É alimentado por modelos de linguagem de grande porte que são pré-treinados em vastos conjuntos de dados e comumente chamados de modelos fundacionais (FM).

Funcionamento dos modelos fundacionais



Treinamento

Inferência

generation

mmarization

tion extraction

Q&A

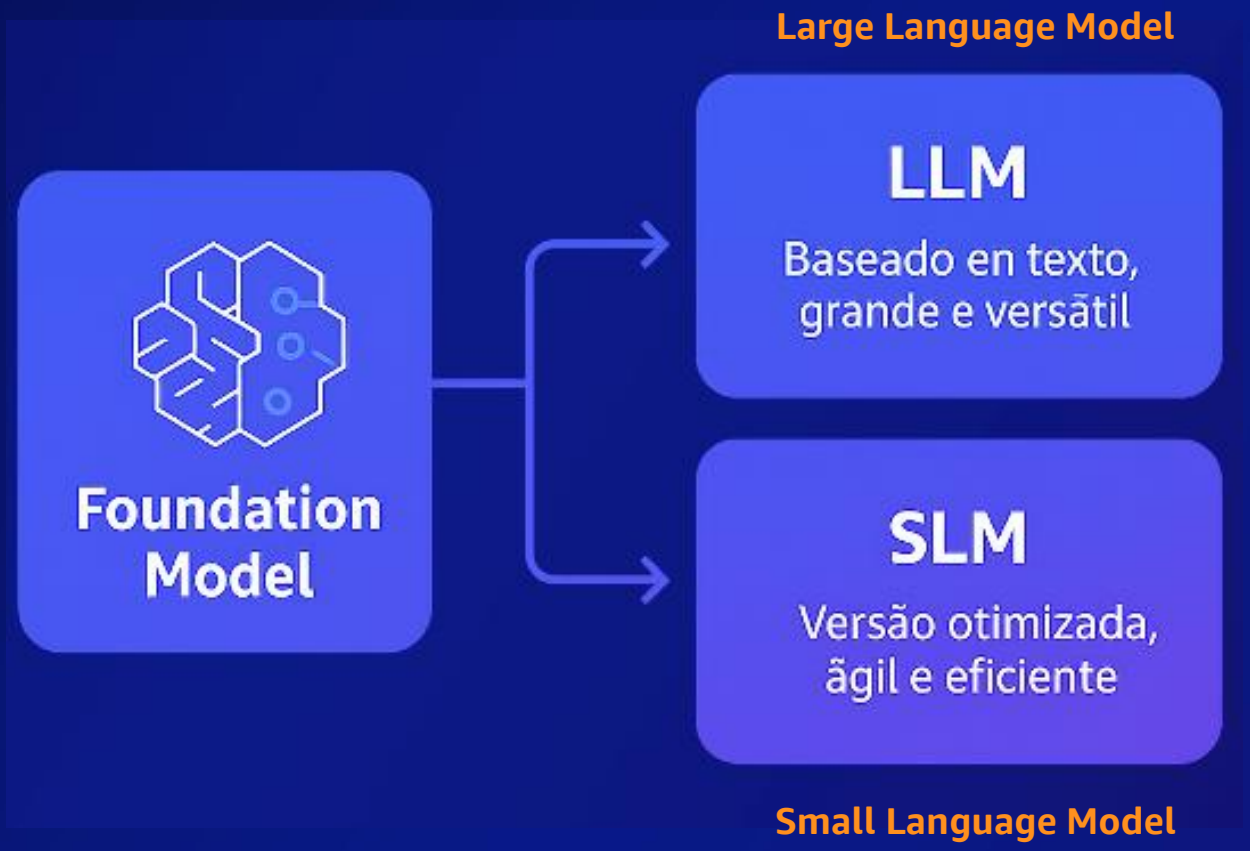
Chatbot

Unlabeled data

FM

Broad range of general tasks

Modelos Fundacionais:
LLMs e SLMs, quais são as diferenças?



Llama 3.1 405B parâmetros

Llama 3.2 3B parâmetros

Quando Usar LLMs e Quando Eles se Tornam um Desafio?

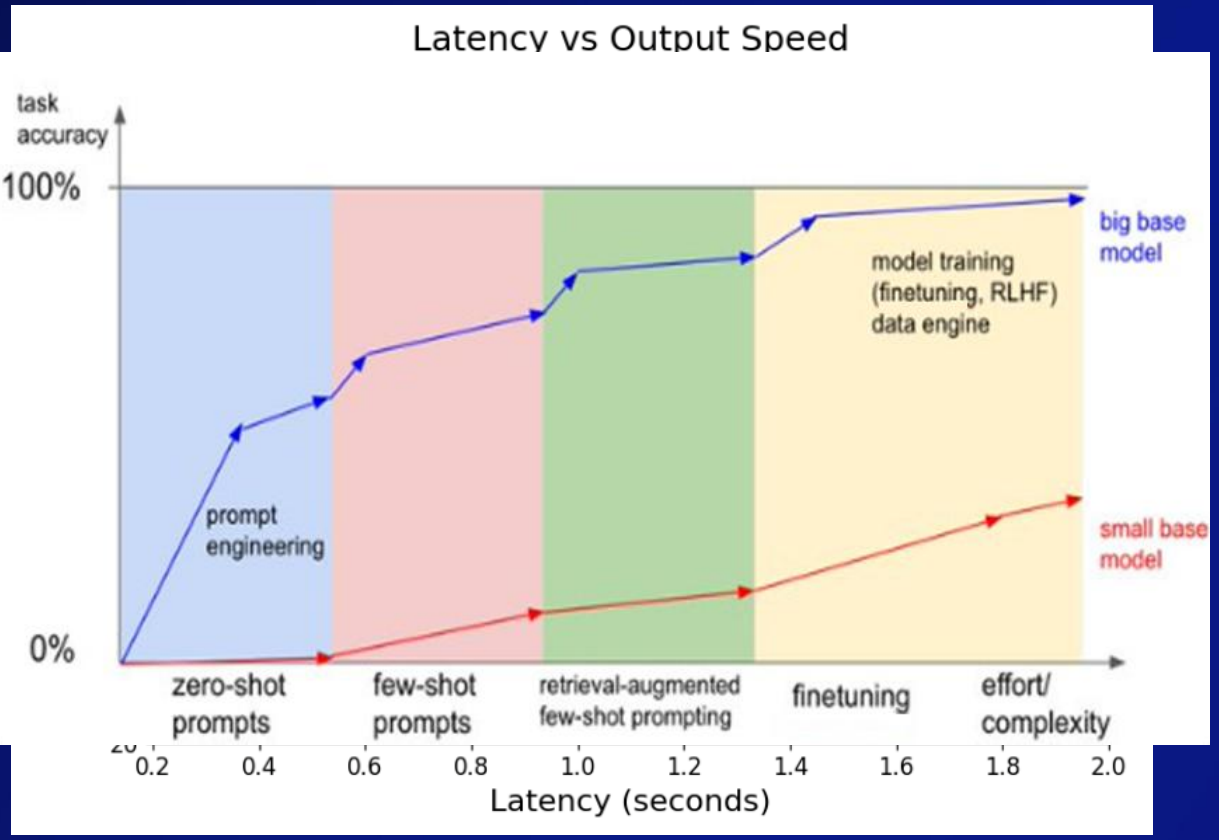
Quando usar LLMs completos (grandes modelos)?

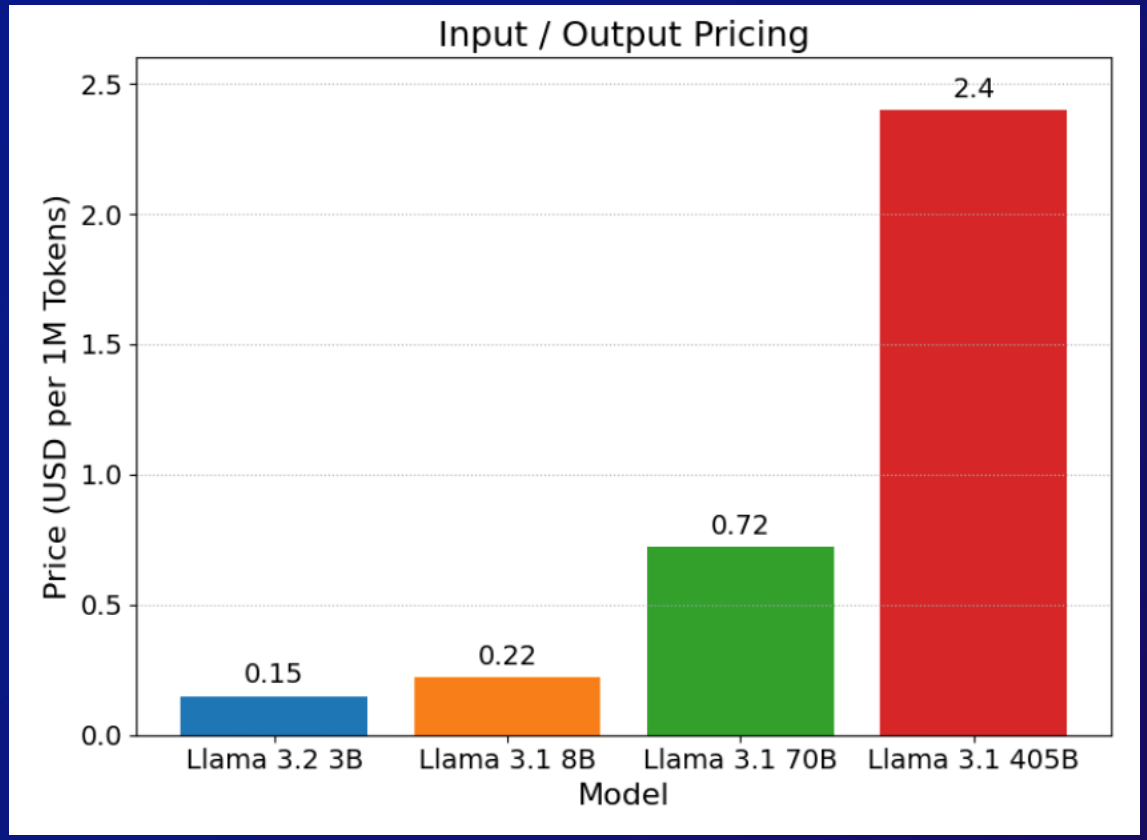
- Quando o problema é **aberto, não estruturado ou criativo**
- Quando a base de conhecimento **muda com frequência**
- Quando **não há histórico de dados rotulados** para treinamento
- Quando a IA precisa **raciocinar com contexto amplo ou múltiplas fontes**

... mas quando se tornam um desafio?

- Em tarefas **repetitivas e previsíveis** (ex: religação, diagnósticos conhecidos)
- Quando o volume de requisições é **muito alto**
- Quando há **latência crítica** (tempo real, edge)
- Quando os custos operacionais **precisam ser controlados com precisão**

Comparando os LLMs e SLMs





Oportunidades e Dilemas da IA Generativa nas Utilities

Desafio

Atendimento Lento

Alarmes Ignorados

Aplicação GenAI

IA gerando boas respostas

Novo Desafio

Alto custo

Atraso

E se tivéssemos um jeito de...

Aproveitar toda a **inteligência** de um LLM grande...

... mas com a **rapidez e o custo** de um modelo pequeno?

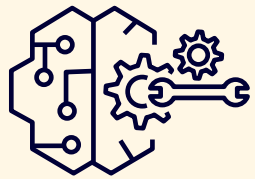
... e ainda garantir que ele funcione **offline** quando necessário?

... e se ele fosse capaz de entender nossos processos técnicos e regras de compliance?

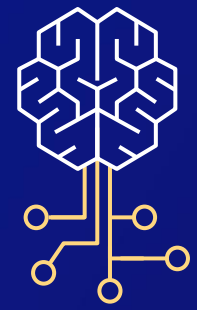
É aqui que entra o uso estratégico de modelos destilados

O que é destilação e como funciona

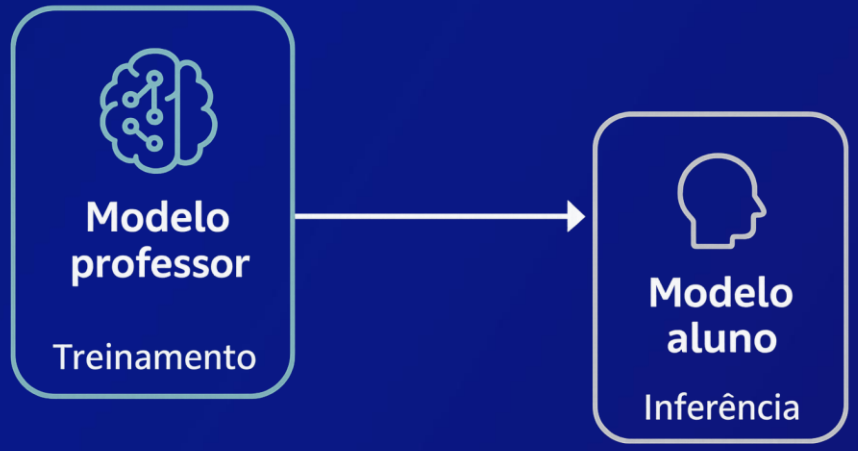
Prompts



Plataforma
Destilação Modelos

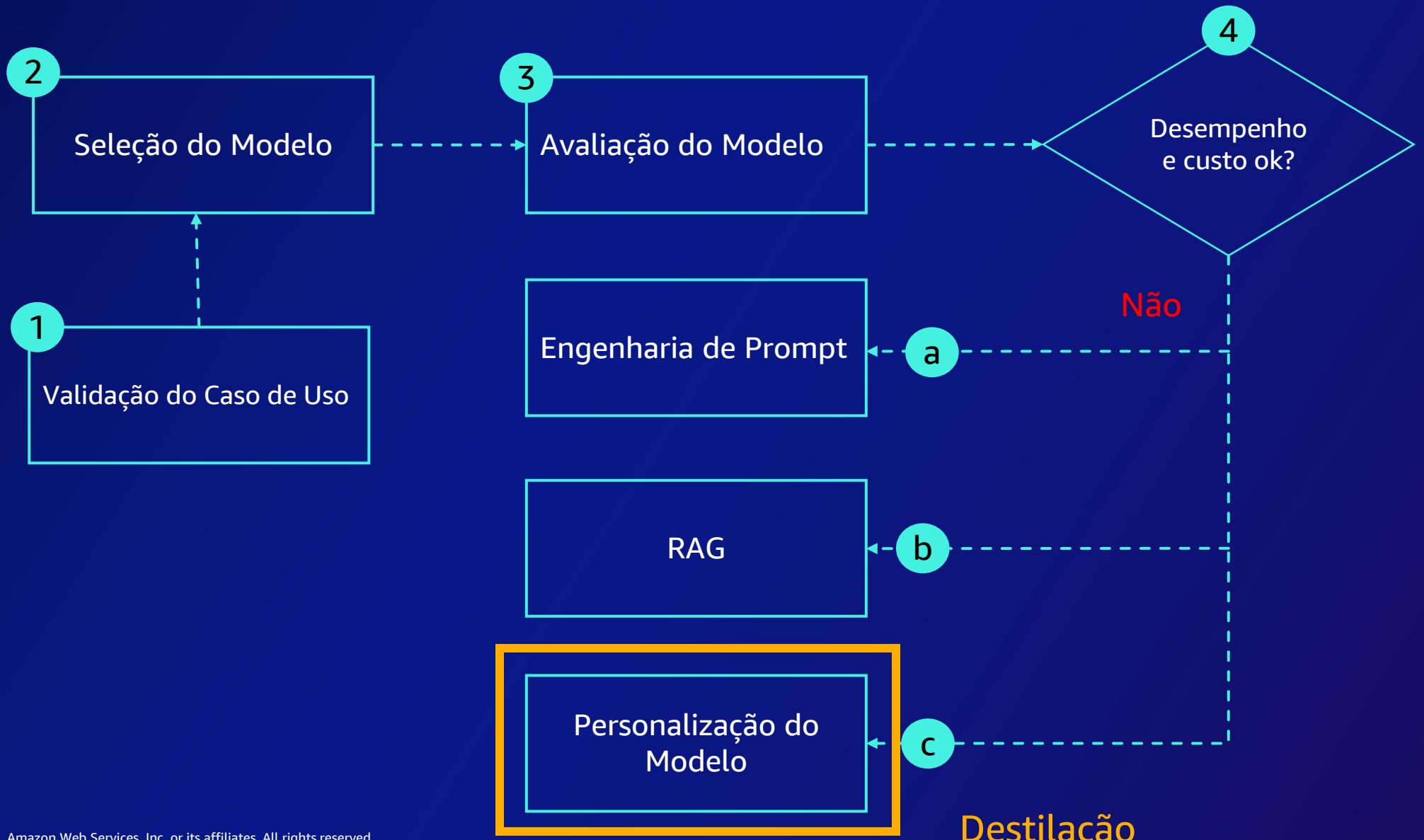


Modelo
Destilado



Desacoplamento

Fluxo de Desenvolvimento de IA Generativa



Stack de IA Generativa da AWS

APLICAÇÕES PARA AUMENTAR A PRODUTIVIDADE

Amazon Q Business
Amazon Q in QuickSight
INSIGHTS AND AUTOMATION

Amazon Q Developer
SOFTWARE DEVELOPMENT LIFECYCLE

MODELOS E FERRAMENTAS PARA CONSTRUIR APLICATIVOS DE IA GENERATIVA

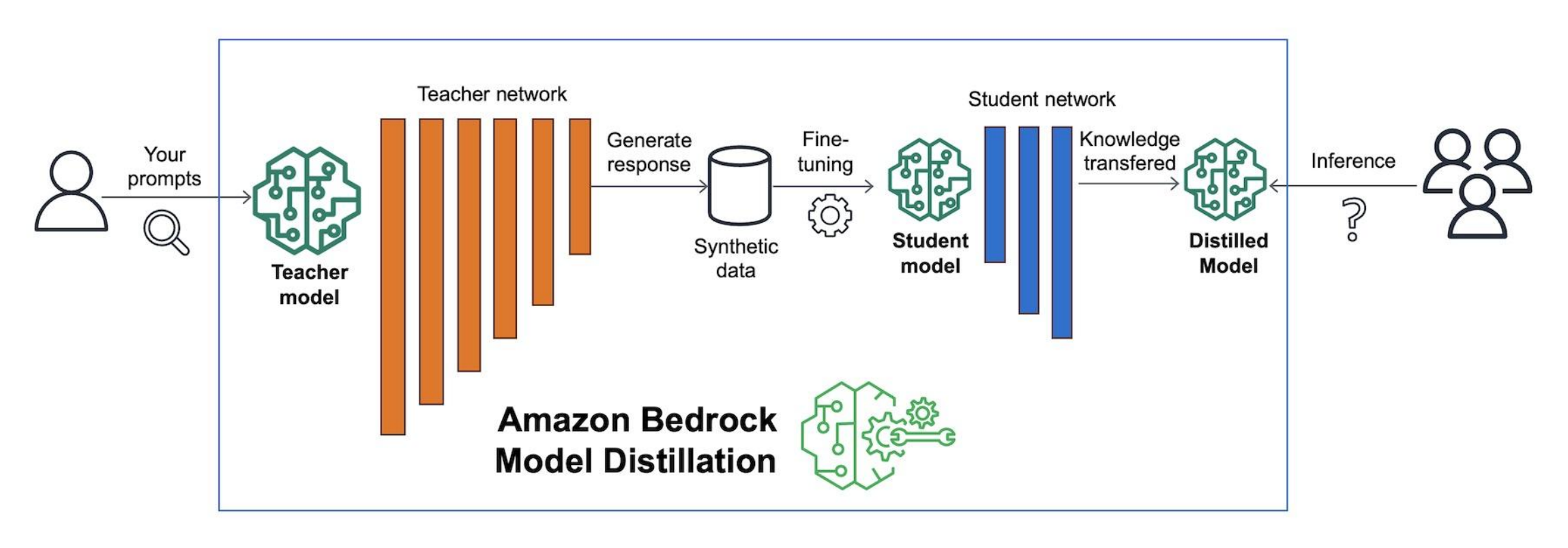
Amazon Bedrock
AMAZON MODELS | PARTNER MODELS

INFRAESTRUTURA PARA CONSTRUIR E TREINAR MODELOS DE IA

Amazon SageMaker
MANAGED INFRASTRUCTURE

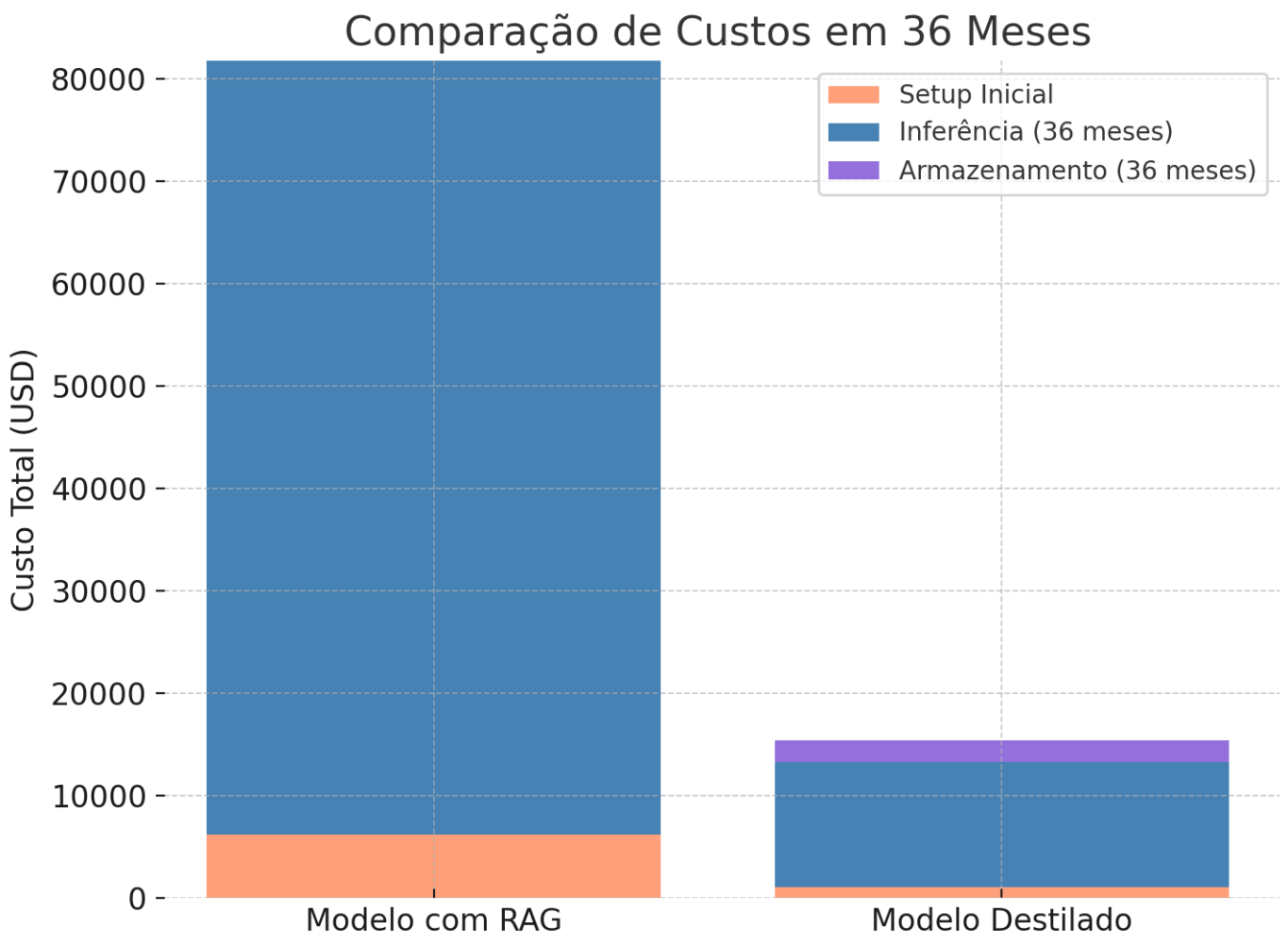
AWS Trainium
AWS Inferentia
GPUs
HIGH PERFORMANCE COMPUTE

Processo de Destilação de Modelos



https://aws.amazon.com/blogs/machine-learning/amazon-bedrock-model-distillation-boost-function-calling-accuracy-while-reducing-cost-and-latency/

Comparativo entre as abordagens



Volume mensal: 1 milhão de inferências.
Tokens de entrada:

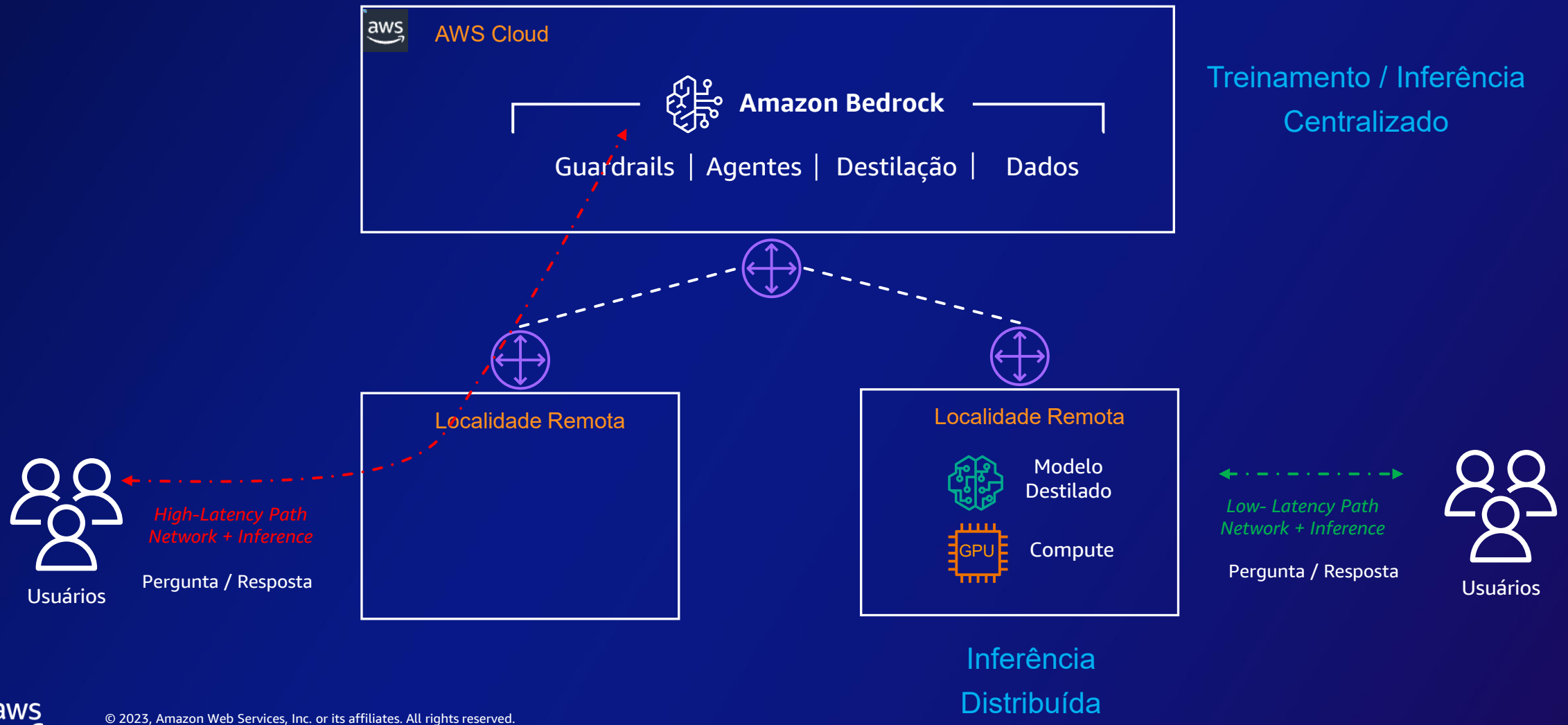
- 200 tokens (query) + 500 tokens (contexto) para modelo com RAG;
- 200 tokens (query) para destilado.

Tokens de saída: 300 tokens para ambos.
Treinamento (destilação): 50 milhões de tokens processados.
RAG: $200-400/mês.
Inferência:

- Nova Premier = $0.00164
- Nova Pro = $0.00028

*Os valores utilizados são públicos e baseados na região **us-east-1** da AWS, apenas para referência.*

Estendendo a Nuvem para uma Experiência Híbrida no Setor de Utilities



Key Takeaways

01

Comece pelo caso de uso

02

Destilação como aliada

03

Uso Estratégico de Agentes de IA

04

Capacitação, Treinamento e Experimentação

Thank you!