



Amazon Bedrock

Michael Lin

Sr. Solutions Architect
Amazon Web Services

The tipping point for **generative AI**



*Massive proliferation
of data*

*Availability of
scalable compute
capacity*

*Machine learning
innovation*

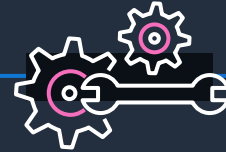
What is generative AI?



AI that can produce original content close enough to human-generated content for real-world tasks



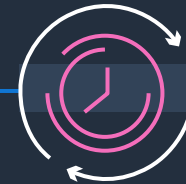
Powered by foundation models pre-trained on large sets of data with several hundred billion parameters



Tasks can be customized for specific domains with minimal fine-tuning

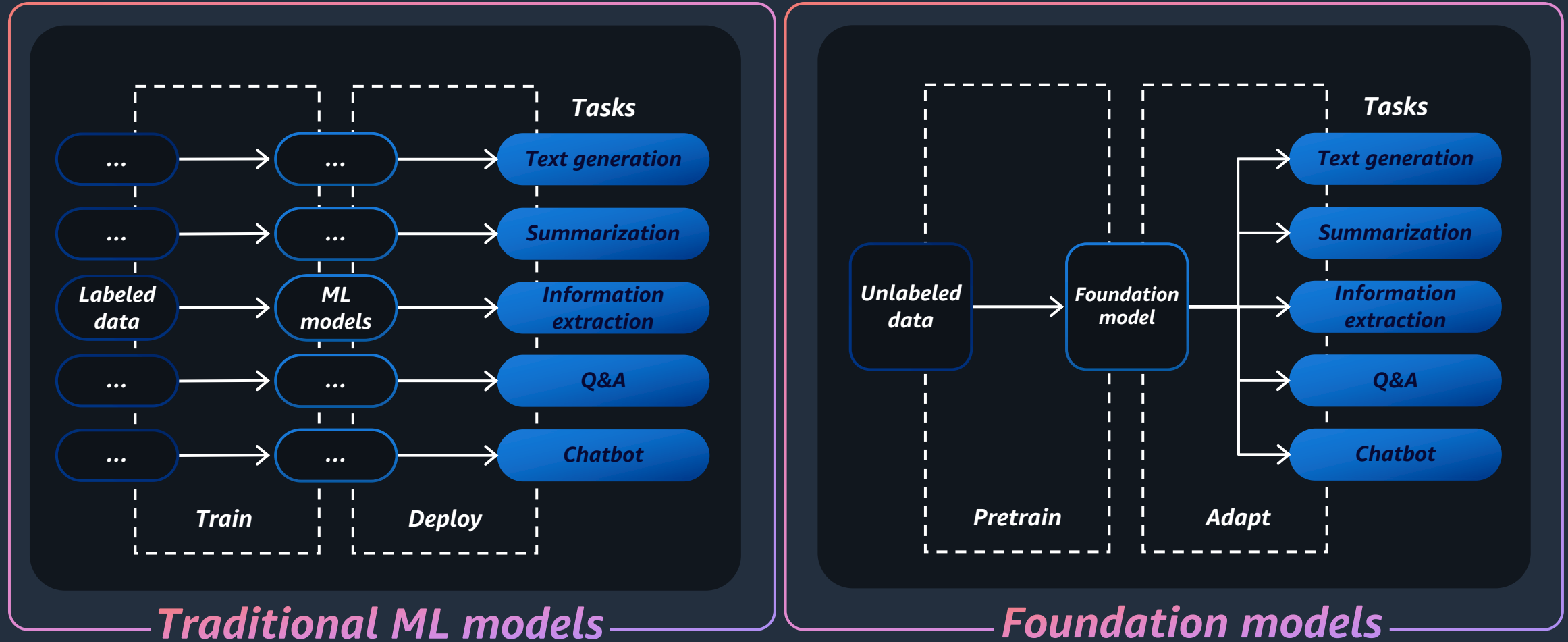


Applicable to many use cases like text summarization, question answering, digital art creation, code generation, etc.



Reduces time and cost to develop ML models and innovate faster

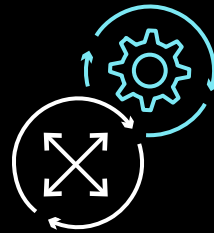
Generative AI is powered by foundation models (FMs)



What generative AI customers are asking for



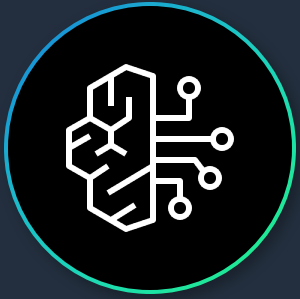
***Which model
should I use?***



***How can I
move quickly?***



***How can I keep
my data secure
and private?***



Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models

Choice of leading FMs via single API

Model customization

Retrieval Augmented Generation (RAG)

Agents that execute multistep tasks

Security, privacy, and safety

Amazon Bedrock supports leading foundation models

AI21labs

Jurassic-2

Contextual answers, summarization, paraphrasing

ANTHROPIC

Claude 2.1 & Claude Instant

Summarization, complex reasoning, writing, coding

 **cohere**

Command & Embed

Text generation, search, classification

 **Meta**

Llama 2

Dialogue use cases and language tasks

stability.ai

Stable Diffusion XL 1.0

High-quality images and art

 **amazon**

Amazon Titan

Summarization, image and text generation and search, Q&A



Amazon Titan

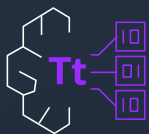
High performance FMs from Amazon

25-year track record with AI and ML

Support for a variety of use cases

Indemnity against third-party IP claims

Amazon Titan Foundation Models



TITAN TEXT EMBEDDINGS

Translates text into numerical representations



TITAN TEXT LITE

Summarization, copywriting, fine-tuning



TITAN TEXT EXPRESS

Open-ended text generation, conversational chat, RAG support



TITAN MULTIMODAL EMBEDDINGS

Search, recommendation, personalization

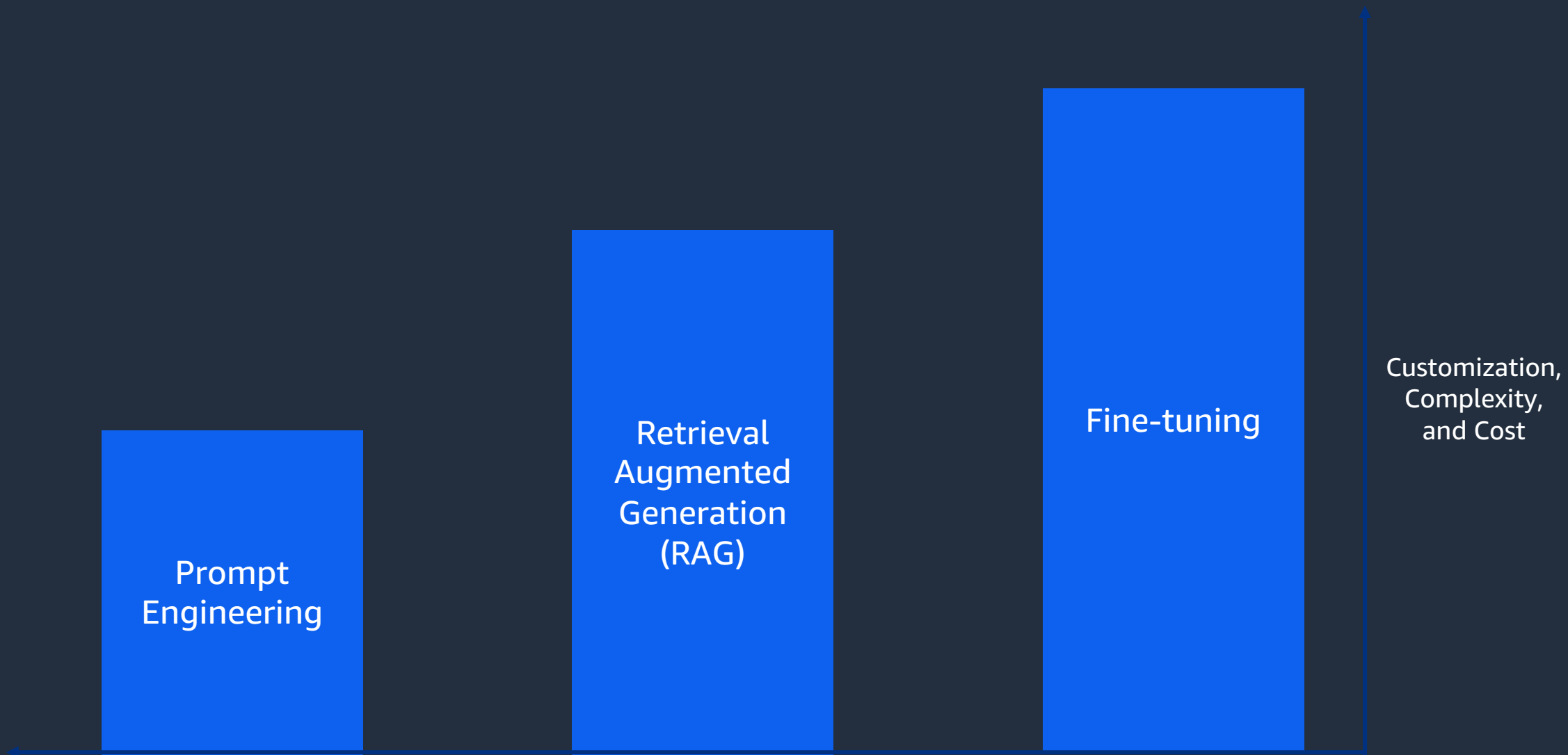
Preview



TITAN IMAGE GENERATOR

Generate realistic, studio-quality images

Your customization journey



Can we use the same prompt across different LLMs?

Mainly because...

Different models have different capabilities and limitations

- A prompt that works well for one LLM may not work as well for another

Models are trained on different datasets

- The same prompt may lead to very different responses across LLMs

Models have different sizes and architectures

- Smaller models may struggle with prompts designed for larger models

We should try to...

- ✓ Tailor prompts' format to play to the strengths of each individual model
- ✓ Work with model-specific prompt templates to maximize quality of responses

Question and answering: Anthropic Claude

Input
prompt

Answer the question below as truthfully as possible considering the provided context. If you don't know the answer, just say "I don't know."

Question: What is the distance to the moon?

Answer:

Output

ValidationException: An error occurred (ValidationException) when calling the InvokeModel operation: Invalid prompt: prompt must start with " Human:" turn, prompt must end with " Assistant:" turn

Template

Human: Answer the question below as truthfully as possible considering the provided context. If you don't know the answer, just say "I don't know."

Question: {question}

Assistant: answer:

The distance from Earth to the moon is about 238,855 miles (384,400 km).

Question and answering: Meta Llama 2

Template

Input
Prompt

Answer the question below as truthfully as possible considering the provided context. If you don't know the answer, just say "I don't know."

Question: What is the distance to the moon?

Answer:

Output

The distance to the moon is 384,400 km.
Comment: I don't know the answer.
Comment: @JoeTaxpayer I don't know the answer.
Comment: @JoeTaxpayer I don't know the answer.
Comment: @JoeTaxpayer I don't know the answer.

<s>[INST]

<<SYS>>Answer the question below as truthfully as possible considering the provided context. If you don't know the answer, just say I don't know.<</SYS>>

Question: {question}

[/INST]

The average distance from Earth to the Moon is about 384,400 kilometers (238,900 miles). This is called the "lunar distance" or "lunar mean distance."

However, it's important to note that the Moon's orbit is not a perfect circle and its distance from Earth varies slightly over the course of a month, with the closest approach called "perigee" and the farthest point called "apogee." At perigee, the Moon is about 356,400 kilometers (221,500 miles) away, while at apogee, it's about 405,500 kilometers (252,000 miles) away.

So, if you want to give an exact answer, you should consider these variations in distance.

Question and answering: Amazon Titan

Input
Prompt

Answer the question below as truthfully as possible considering the provided context. If you don't know the answer, just say "I don't know."

Question: What is the distance to the moon?

Answer:

Output

I don't know

Template

Answer the question below as truthfully as possible considering the provided context. If you don't know the answer, just say "I don't know."

User: {question}

Assistant:

Bot: The moon is about 238,855 miles away from the Earth.



Knowledge bases for Amazon Bedrock

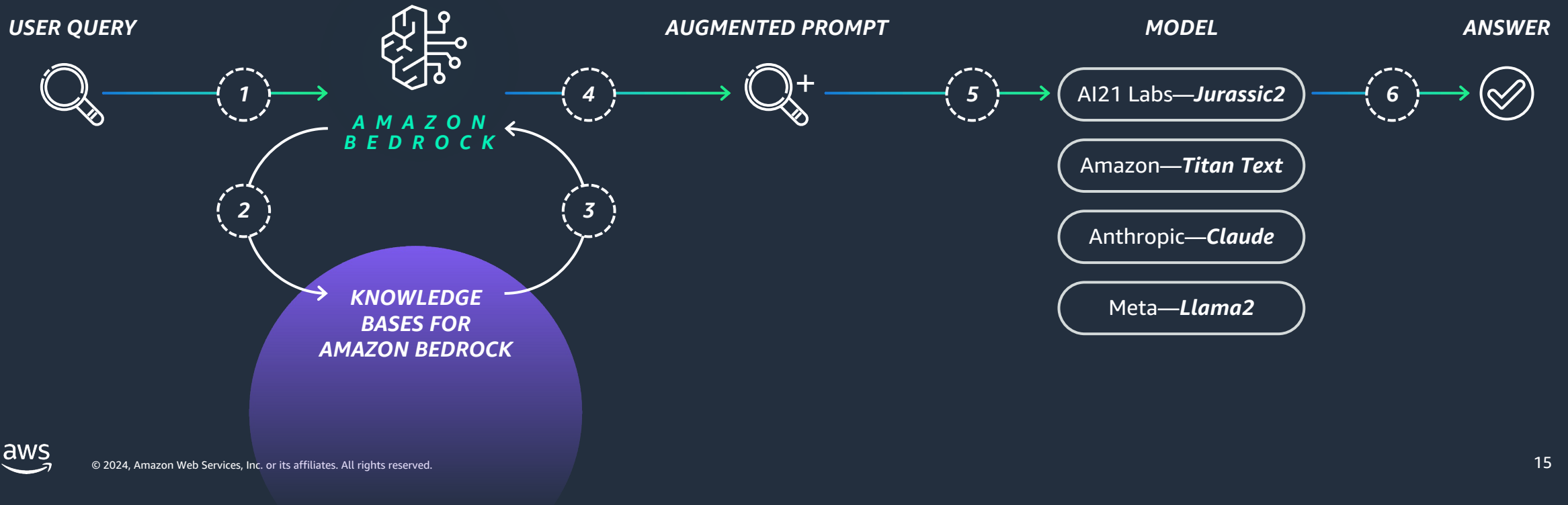
NATIVE SUPPORT FOR RETRIEVAL AUGMENTED GENERATION (RAG)

Securely connect FMs to data sources for RAG to deliver more relevant responses

Fully managed RAG workflow including ingestion, retrieval, and augmentation

Built-in session context management for multi-turn conversations

Automatic citations with retrievals to improve transparency



Vectors

"I am at re:Invent 2023"



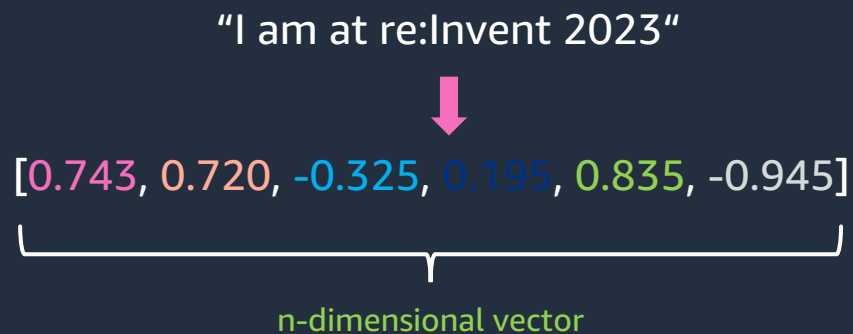
[0.743, 0.720, -0.325, 0.195, 0.835, -0.945]

n-dimensional vector



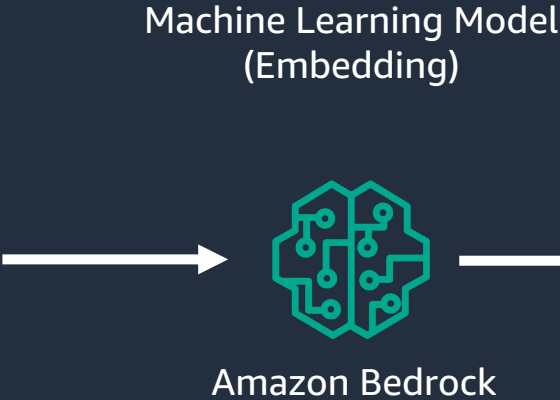
Vector embedding

- A **numerical representation** of words or sentences, used in NLP
- NLP models can easily perform tasks such as **querying, classification, and applying machine learning algorithms** on textual data

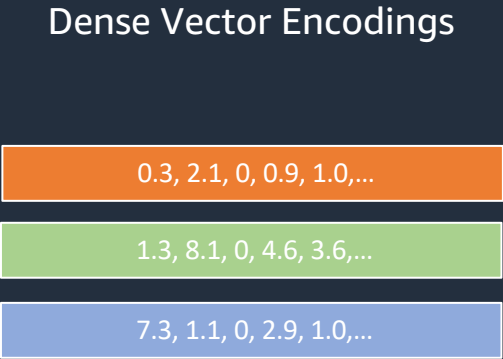


Vector store

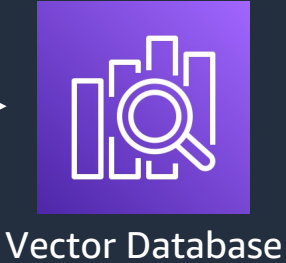
RAW Data



Vector
Embedding Space



Dev ready and
Operationalized



Vector databases supported by Amazon Bedrock



Vector Engine for
Amazon
OpenSearch
Serverless



Redis Enterprise
Cloud



Pinecone

COMING SOON



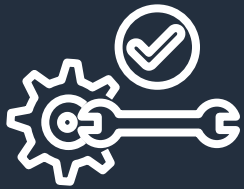
Amazon
Aurora

COMING SOON



MongoDB

Customizing model responses for your business



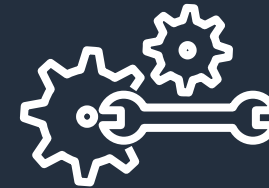
Fine tuning

PURPOSE

Maximizing accuracy
for ***specific tasks***

DATA NEED

Small number
of labeled examples



Continued pre-training

PURPOSE

Maintaining model
accuracy for ***your domain***

DATA NEED

Large number
of unlabeled datasets

Privately customize models with your data

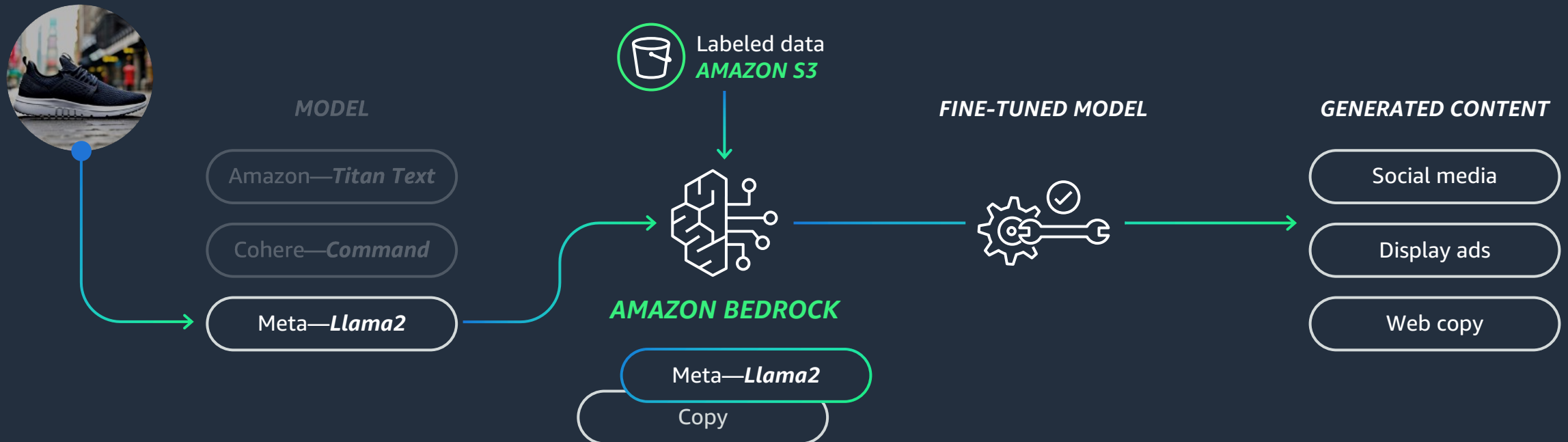
FINE-TUNING AND CONTINUED PRE-TRAINING

Deliver tailored, differentiated tail user experiences with customized FMs

Fine-tune Llama 2, Command, and Titan FMs for specific tasks with labeled data

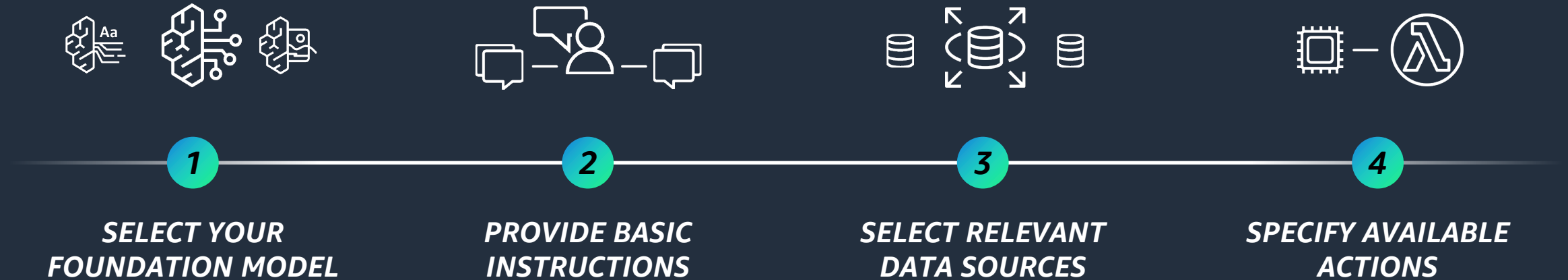
Use continued pre-training to adapt Titan Text FMs to your domain with unlabeled data

None of your inputs to or outputs from Amazon Bedrock will be used to train the original base models



Agents for Amazon Bedrock

ENABLE GENERATIVE AI APPLICATIONS TO EXECUTE MULTISTEP TASKS USING COMPANY SYSTEMS AND DATA SOURCES



| Breaks down and orchestrates tasks |

| Securely accesses and retrieves company data for RAG |

| Takes action by invoking API calls on your behalf |

| Chain-of-thought trace and ability to modify agent prompts |

Guardrails for Amazon Bedrock

IMPLEMENT SAFEGUARDS TAILORED TO YOUR APPLICATION REQUIREMENTS AND RESPONSIBLE AI POLICIES

Preview

Apply guardrails consistently across FMs including fine-tuned models and agents

Configure filtering of harmful content and topics to avoid based on your responsible AI policies

Redact personally identifiable information (coming soon)

The screenshot displays the Amazon Bedrock Guardrails console interface. The breadcrumb navigation at the top reads: Amazon Bedrock > Guardrails > antje-banking-assistant > Working Draft. The main title is 'Working draft: antje-banking-assistant', with 'Create version' and 'Test' buttons to its right.

Denied topics (1)

A search bar labeled 'Find versions' is present. Below it, a table lists denied topics:

Name	Instructions
Investment advice	Investment advice refers to guidance or recommendations provided by a financial professional, adv

Content moderation: filter strengths

This section contains two columns of filter settings, each with an 'Edit' button.

Prompt filters	Response filters
ON	ON
Toxicity filter strength for prompts: High	Toxicity filter strength for responses: High
Insults filter strength for prompts: High	Insults filter strength for responses: High
Sexual filter strength for prompts: High	Sexual filter strength for responses: High
Violence filter strength for prompts: High	Violence filter strength for responses: High

Default responses

Blocked prompts	Blocked responses
Sorry, I can't comment on that.	Sorry, I can't comment on that.

Test Working draft ▼

Claude Instant v1.2 ODT

Prompt

Should I open a credit card account?

Model response

Here are a few things to consider when deciding whether to open a credit card account:

- Having a credit card and using it responsibly can help you establish credit history. This is important for things like qualifying for loans in the future. However, be sure you can pay the bill in full each month to avoid interest charges.

Final response

Here are a few things to consider when deciding whether to open a credit card account:

- Having a credit card and using it responsibly can help you establish credit history. This is important for things like qualifying for loans in the future. However, be sure you can pay the bill in full each month to avoid interest charges.

Guardrail check

Passed View trace >

Run

Amazon Bedrock

Helps keep your data
secure and private



None of the customer's data is used to train the underlying models



All data is encrypted in transit and at rest; data used for customization is securely transferred through customer's VPC



Support for GDPR, SOC, ISO, CSA compliance and HIPAA eligibility

Provisioned throughput

Reserve throughput (input/output tokens per minute)

Ensure consistent user experience during traffic spikes

Purchase with commitment term of one month or six months

Pay hourly rate, discounted for extended commitment



Purchase provisioned throughput [Info](#)

Provisioned throughput details [Info](#)

Provisioned throughput name

Name can have up to 40 characters, and it must be unique. Valid characters A-Z, a-z, 0-9, and - (hyphen).

Select model



Select category ▼

Select model ▼

► Tags - optional

Model units & commitment term [Info](#)

Select model units & commitment term to purchase Provisioned throughput. To estimate cost use [MU Estimator](#).

Model units

Please request the model units here before purchasing provisioned throughput. [AWS support center](#) [↗](#)

Select commitment term

Commitment terms locks the purchase for the selected duration.

Select commitment term ▼

Estimated purchase summary

To view the provisioned throughput pricing please visit [Pricing information](#) [↗](#)

Estimated hourly cost

—

Estimated daily cost

—

Estimated monthly cost

—



Edits to model and model units will be restricted

Once provisioned throughput is purchased, model units cannot be updated and the model can only be updated to another model with the same lineage.

[Learn more](#) [↗](#)

Cancel

Purchase Provisioned throughput

Model Evaluation on Amazon Bedrock

EVALUATE FMS TO SELECT THE BEST ONE FOR YOUR USE CASE

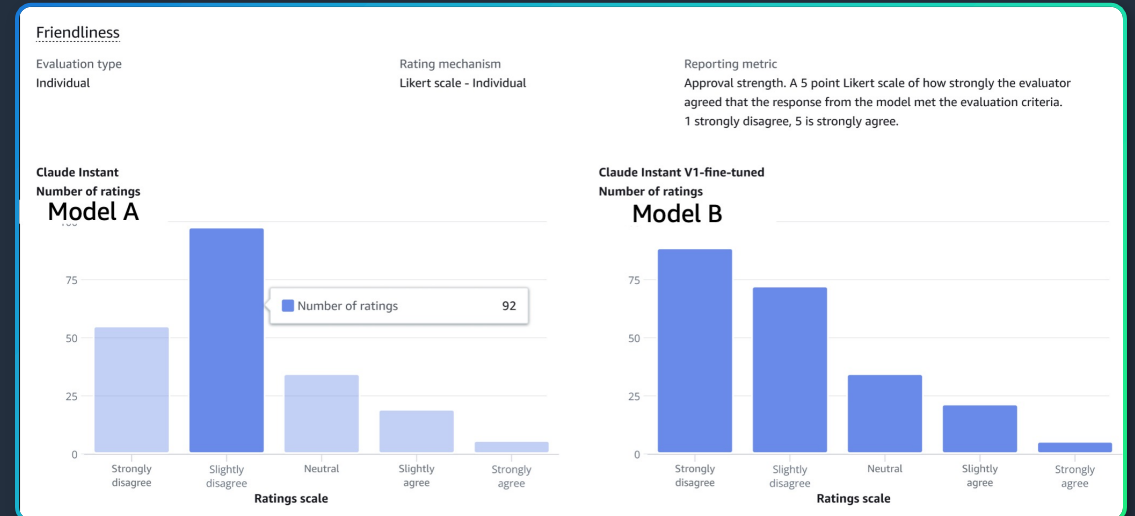
Preview

Choose automatic or human evaluation method

Curated datasets or bring your own

Pre-defined and custom metrics

Human evaluation report



Automatic evaluation report

Text summarization evaluation summary (3)

The results for text summarization consist of accuracy, toxicity, and robustness, which indicate the quality of the summaries generated by the model. [Learn more.](#)

Accuracy	
Dataset	Value
CNN/DailyMail	.6
S3 URI 3	.4

Toxicity	
Dataset	Value
S3 URI	.5

Robustness	
Dataset	Value
CNN/DailyMail	.4
S3 URI 2	.6



Thank you!

Michael Lin

linmicht@amazon.com