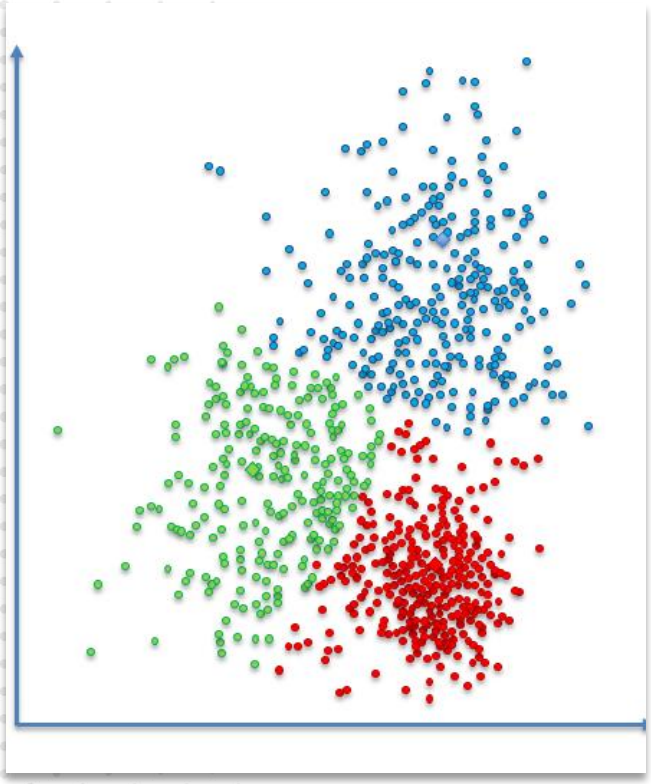


Lesson #5:
Cloud-Based LLM Development Platforms

Objectives:

- Overview of Cloud Platforms
- Enter AWS Bedrock
- Foundation Model Providers and Access
- Anatomy of a Bedrock Call
- Going Deeper

Cloud Development Platforms

Models and Transformers
- Llama
- OpenAI, Claude, etc
- DeepSeek

Agent Orchestration Platforms
- LangGraph
- AutoGen
- CrewAI

Development Environments
- Colab
- VS Code
- Cursor / Windsurf

Agentic Protocols
- APIs
- MCP
- A2A

Local LLM Management Platforms
- Ollama
- LM Studio
- Jan

Cloud Development Platforms
- AWS Bedrock
- Azure AI Studio
- Google Vertex AI

RAG & Agentic Frameworks
- LlamaIndex
- LangGraph
- Haystack

Monitoring and Observability
- LangSmith
- Weights & Biases
- Helicone

Mature | Strong Contenders | Emerging

Pearson

Cloud LLM Platforms

- AWS, Google, and Azure support LLM model development in a cloud setting.
- They do the following:
  - Provide Access to Foundation Models
  - Offer Serverless Inference APIs
  - Support Model Customization
  - Enable Retrieval-Augmented Generation (RAG)
  - Provide Agentic & Tool-Using Frameworks
  - Support Evaluation, Monitoring & Guardrails

The Big Players

- AWS Bedrock
  - Serverless API access to top foundation models (Claude, Mistral, LLaMA, Titan)
- Azure AI Studio
  - Build AI copilots using GPT-4 etc., with deep integration into Office 365, security, and compliance controls.
- Google Vertex
  - Suited to advanced AI engineers / powerful tools to train/tune/customize LLMs.

Comparison

	AWS Bedrock	Azure AI Studio	Google Vertex
Overview	Fully managed GenAI platform for FMs	Unified AI development platform for Azure OpenAI & more	Full-service MLOps & GenAI platform
Model access	Anthropic (Claude), Meta (LLaMA), Mistral, Cohere, etc.	OpenAI (GPT), Meta (LLaMA), Mistral, etc.	Gemini, PaLM, LLaMA, Mistral, etc.
Serverless inference	Yes	Yes	Yes
Agent framework	Bedrock Agents (API orchestration + tools), Sagemaker integration	Manual tool calling with functions	Agent Builder for tool use, planner, memory
Suited for	Teams needing fast access to multiple hosted FMs	Enterprises with Microsoft stack, Azure OpenAI users	GCP users and advanced AI/ML engineer

Introduction to AWS Bedrock

aws | Search [Option+S] | United States (Ohio)

Amazon Bedrock

Amazon Bedrock

- **Getting started**
  - Overview
  - Providers
- **Foundation models**
  - Model catalog New
  - Marketplace deployments New
  - Prompt Routers
- **Playgrounds**
  - Chat / Text
- **Builder tools**
  - Agents
  - Flows
  - Knowledge Bases
  - Prompt Management
- **Safeguards**
  - Guardrails
- **Inference and Assessment**
  - Provisioned Throughput
  - Batch inference
  - Cross-region inference
  - Evaluations

Introducing Prompt routers
Route requests between foundational models from the same family, optimizing for response quality and cost.
View Prompt routers

Overview Info

Foundation models

Amazon Bedrock supports over 100 foundation models from industry-leading providers and emerging leaders. Select a serverless model or Bedrock Marketplace model that is best suited for achieving your unique goals.

View Model catalog | Discover marketplace models

Model spotlight

Anthropic's Clauc

Choose the exact combination of intelligence, speed, and cost to suit y needs. All of the latest Claude mode Claude 4, are available in Amazon Be

Open in chat playground

Chat / Text

Generate text for a vast range of language processing tasks with various pre-trained models. You can use a single prompt or iterate on the result by submitting subsequent prompts that take into acco context of previous prompts and generated responses in a chat format.

Open playground

Builder tools

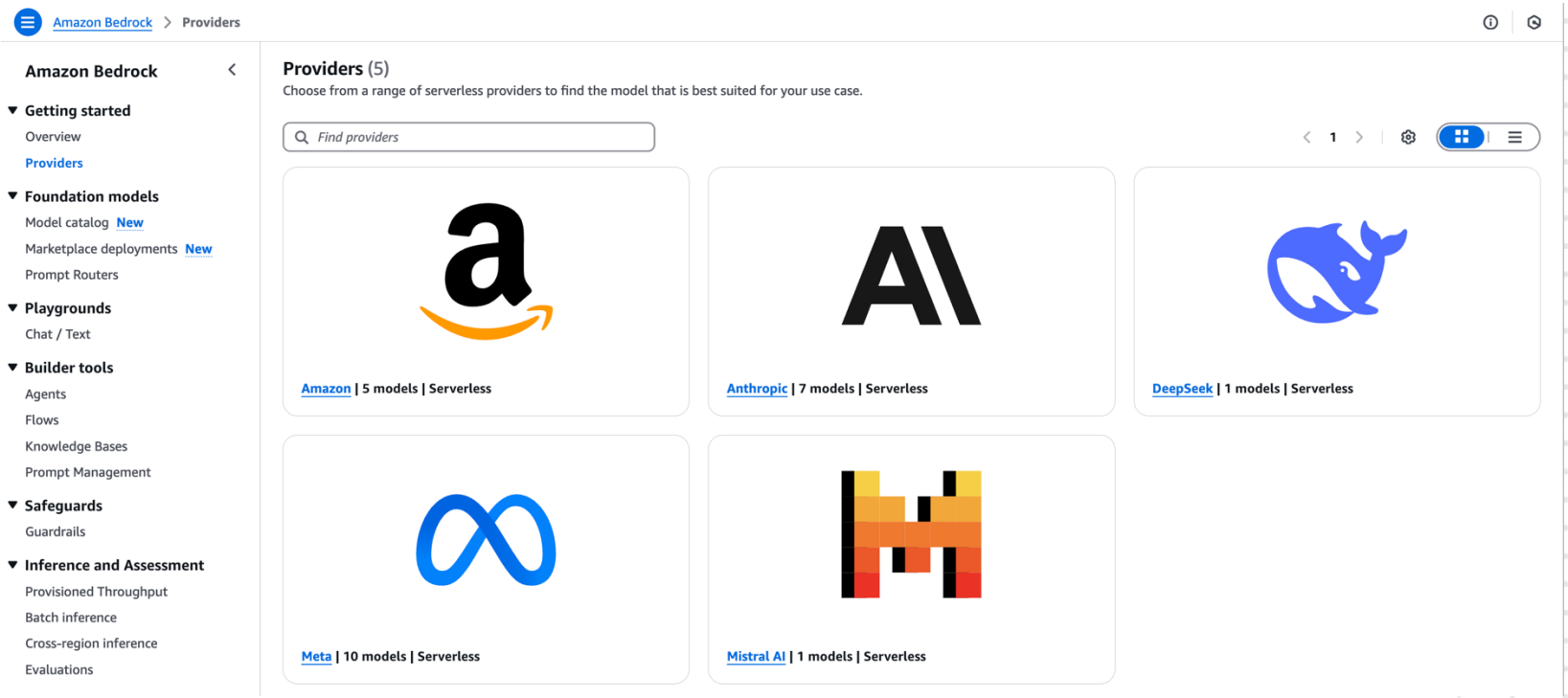
Create GenAI applications, augment responses with your proprietary data, and experiement with prompts.

Region Availability

- Primarily us-east-1 and us-west-2 (most Bedrock services are available)
- There is variance by model provider – not all models are available everywhere
- What do I do if I'm in a different region?
  - Some models are accessible through cross-Region inference

- Refer here for updated region support: https://docs.aws.amazon.com/bedrock/latest/userguide/models-regions.html

Who are the Model Providers?



Model Access

aws | AWS Console Home | Search [Option+S] | United States (Ohio)

Amazon Bedrock > Model access > Manage model access

Step 1
Edit model access

Step 2
Review and submit

Edit model access

Base models (3/24)

Collap

Not seeing a model you're interested in? Check out all supported models by region here.

Find model

Group by provider

Models	Access status	Modality	EULA
Amazon (5)	0/5 access granted		
Titan Text Embeddings V2	Available to request	Embedding	EULA
Nova Pro Cross-region inference	Available to request	Text & Vision	EULA
Nova Lite Cross-region inference	Available to request	Text & Vision	EULA
Nova Micro Cross-region inference	Available to request	Text	EULA
Nova Premier Cross-region inference	Available to request	Text & Vision	EULA
Anthropic (7)	1/7 access granted		
Claude 3 Haiku Cross-region inference	Available to request	Text & Vision	EULA
Claude 3.5 Sonnet Cross-region inference	Available to request	Text & Vision	EULA
Claude 3.5 Haiku Cross-region inference	Available to request	Text	EULA
Claude 3.5 Sonnet v2 Cross-region inference	Available to request	Text & Vision	EULA
Claude 3.7 Sonnet Cross-region inference	Available to request	Text & Vision	EULA
Claude Sonnet 4 Cross-region inference	Access granted	Text & Vision	EULA
Claude Opus 4 Cross-region inference	Available to request	Text & Vision	EULA
DeepSeek (1)	0/1 access granted		
DeepSeek-R1 Cross-region inference	Available to request	Text	EULA

Playgrounds

aws | Search [Option+S] | United States (Ohio)

AWS Console Home

Amazon Bedrock > Chat / Text playground

Mode: Chat

Compare mode

Configurations Info

Llama 3.2 3B In... v1
US Meta Llama 3.2 3B Instruct

Randomness and diversity

Temperature 0.5

Top P 0.9

Length

Response length 512

Guardrails

Manage guardrails

Reset all to default

Input: 43 Output: 407 Latency: 5928 ms

What is the James Webb Space Telescope?

The James Webb Space Telescope (JWST) is a space observatory launched in December 2021 by NASA, the European Space Agency (ESA), and the Canadian Space Agency (CSA). It is the successor to the Hubble Space Telescope and is designed to study the universe in infrared light, providing unprecedented insights into the formation of the universe, the formation of stars and galaxies, and the origins of life.

The JWST is named after James E. Webb, the second administrator of NASA, who played a key role in the Apollo program. The telescope is about the size of a school bus and weighs over 6,000 pounds (2,700 kg). It is equipped with a 6.5-foot (2-meter) diameter primary mirror, made up of 18 hexagonal segments that unfold in space, allowing it to observe the universe in infrared light.

The JWST has several key instruments, including:

1. Near-Infrared Camera (NIRCam): captures high-resolution images of the universe in the near-infrared spectrum.
2. Near-Infrared Spectrograph (NIRSpec): analyzes the light from distant objects, providing information about their composition and motion.
3. Mid-Infrared Instrument (MIRI): observes the universe in the mid-infrared spectrum, allowing it to study cool objects such as planets and asteroids.
4. Fine Guidance Sensor (FGS): helps the telescope track its position and orientation in space.

The JWST has several key science objectives, including:

1. Studying the formation of the first stars and galaxies in the early universe.
2. Observing the formation of planets and the search for biosignatures.
3. Investigating the formation of stars and planetary systems.
4. Understanding the evolution of galaxies and galaxy clusters.

The JWST is designed to operate for at least 5 years, with the potential to extend its mission to 10 years or more. It is a groundbreaking instrument that will revolutionize our understanding of the universe and its many

Write a prompt. Press Shift + Enter to add a new line. Press Enter to generate a response.

Run

Model Hyperparameters

- Temperature: a measure of how creative the model is
  - low Temperature = safe, predictable, deterministic answers. High Temperature = bold, imaginative ones with an increasing amount of creativity.

Temp	Sample Model Response
0.2	The sky appears blue because of Rayleigh scattering. This occurs when sunlight passes through the atmosphere and shorter wavelengths of light, such as blue, are scattered more than longer wavelengths like red.
1	The sky looks blue because sunlight interacts with Earth's atmosphere. As sunlight passes through the air, shorter wavelengths — like blue — scatter more easily in all directions. This scattering, known as Rayleigh scattering, causes us to see the sky as blue most of the time. It's the same reason sunsets can appear red or orange — the angle of the sun changes how much the light is scattered.
2	The sky appears blue due to the scattering of solar radiation by molecules in Earth's atmosphere, particularly a phenomenon called Rayleigh scattering. But if you really think about it, maybe the sky is blue because it's Earth's way of dreaming—its way of whispering ocean memories back to the clouds. After all, light is just nature's poetry in motion.
5	Because marshmallow theories of existential yogurt synchronize with nebula-based fashion regulations. Frogs in bowler hats interpret wavelengths through interpretive dance on Tuesdays, especially near cosmic pancakes. Hence, blue. Obviously.
100	Skibble zentropics cloudmonger flibber astro-tuba syntax bananas! Plasmic shoelaces dangle from hypothetical wombats under bureaucratic sky-harmoniums. Reason? Smeep.

Top-K Sampling

- Instead of selecting a token from the entire vocabulary, the model focuses only on the top K most likely tokens and reassigns probabilities within this smaller group. This introduces controlled randomness.

  Example: if K = 5, the model would narrow its choices to just five tokens:

  - "the" (30%)
  - "a" (25%)
  - "this" (20%)
  - "that" (15%)
  - "one" (10%)

Top-P (nucleus) Sampling

- Similar to Top-K, but instead of being a fixed number, it dynamically adjusts the size of the candidate pool based on a cumulative probability mass

- Example: if P is set to 0.5, it will only look at the tokens that have a cumulative probability of 50%, and will ignore the rest.
  - E.g. if it takes the top 78 tokens to reach 50%, only these will be examined as a next token option.

Connecting to Bedrock Inference Models

- Need to do a few things to prep:
  - Ensure AWS configuration is set up on your machine
  - Set up a python virtual environment
  - Install the boto3 SDK
  - Ensure you know the right AWS region you will connect to

Setting up AWS Access on your Machine

(may need to install it with brew)

```
robbarto@ROBBARTO-M-C6YD .aws % aws configure
AWS Access Key ID [None]: 12323
AWS Secret Access Key [None]: 12321341235
Default region name [None]: us-east-1
Default output format [None]: json
```

```
robbarto@ROBBARTO-M-C6YD .aws % ls -al
total 16
drwxr-xr-x@  4 robbarto  staff   128 Jun 19 12:04 .
drwxr-x---+ 68 robbarto  staff  2176 Jun 19 14:26 ..
-rw-------@  1 robbarto  staff    54 Jun 19 12:04 config
-rw-------@  1 robbarto  staff   117 Jun 19 12:02 credentials
robbarto@ROBBARTO-M-C6YD .aws %
```

Demo Time!

DEMO

1. Confirm you are connecting to AWS Bedrock
2. Perform simple inference to a model
3. Try a Streamlit chatbot with using a Bedrock model