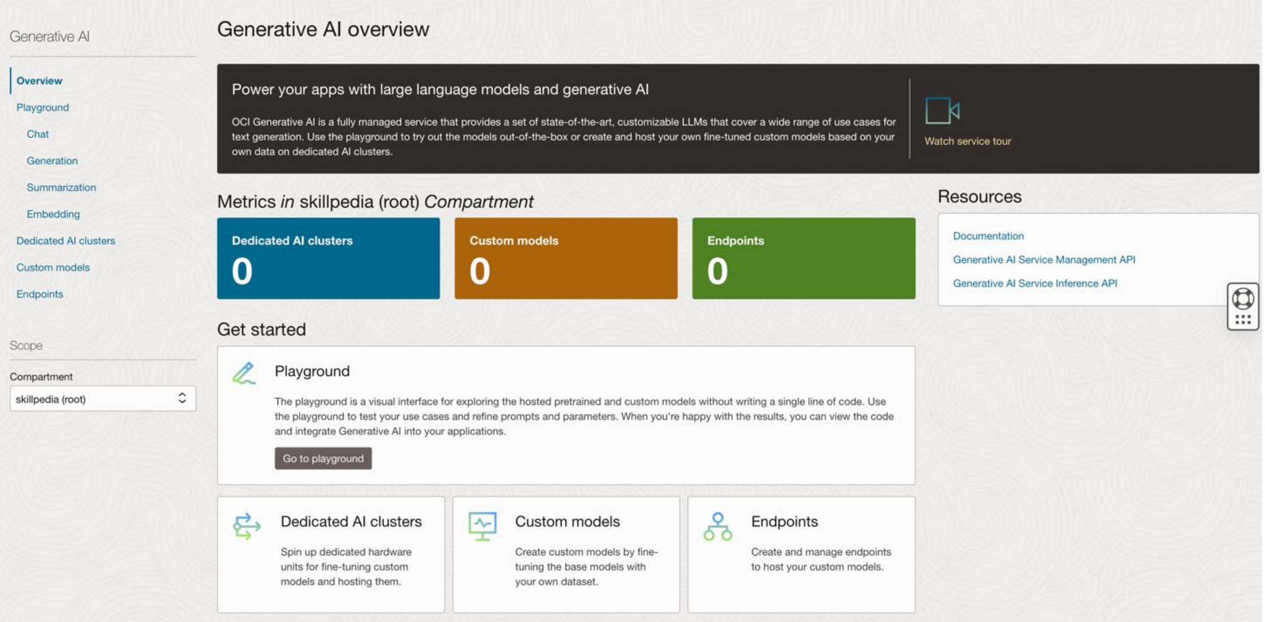
OCI Generative AI - Introduction

Generative AI and Prompt Engineering 2025

Ram N Sangwan

- Introduction to OCI Generative AI Service
- OCI Generative AI API End-Points
- OCI Generative AI Custom Models
- Fine-tuning and Inference in OCI Generative AI
- OCI Generative AI Security

OCI Generative AI Service



Generative AI overview

Power your apps with large language models and generative AI

OCI Generative AI is a fully managed service that provides a set of state-of-the-art, customizable LLMs that cover a wide range of use cases for text generation. Use the playground to try out the models out-of-the-box or create and host your own fine-tuned custom models based on your own data on dedicated AI clusters.

Metrics *in* skillpedia (root) *Compartment*

Dedicated AI clusters	Custom models	Endpoints
0	0	0

Resources

- Documentation
- Generative AI Service Management API
- Generative AI Service Inference API

Get started

Playground

The playground is a visual interface for exploring the hosted pretrained and custom models without writing a single line of code. Use the playground to test your use cases and refine prompts and parameters. When you're happy with the results, you can view the code and integrate Generative AI into your applications.

Dedicated AI clusters

Spin up dedicated hardware units for fine-tuning custom models and hosting them.

Custom models

Create custom models by fine-tuning the base models with your own dataset.

Endpoints

Create and manage endpoints to host your custom models.

Fully managed service that provides a set of customizable LLMs available via a single API to build generative AI applications.

- Choice of Models: high performing pretrained foundational models from Meta and Cohere.
- Flexible Fine-tuning: create custom models by fine-tuning foundational models with your own data set.
- Dedicated AI Clusters: GPU based compute resources that host your fine-tuning and inference workloads.

Pretrained Models (as of 29/6/2025, Chicago Region)

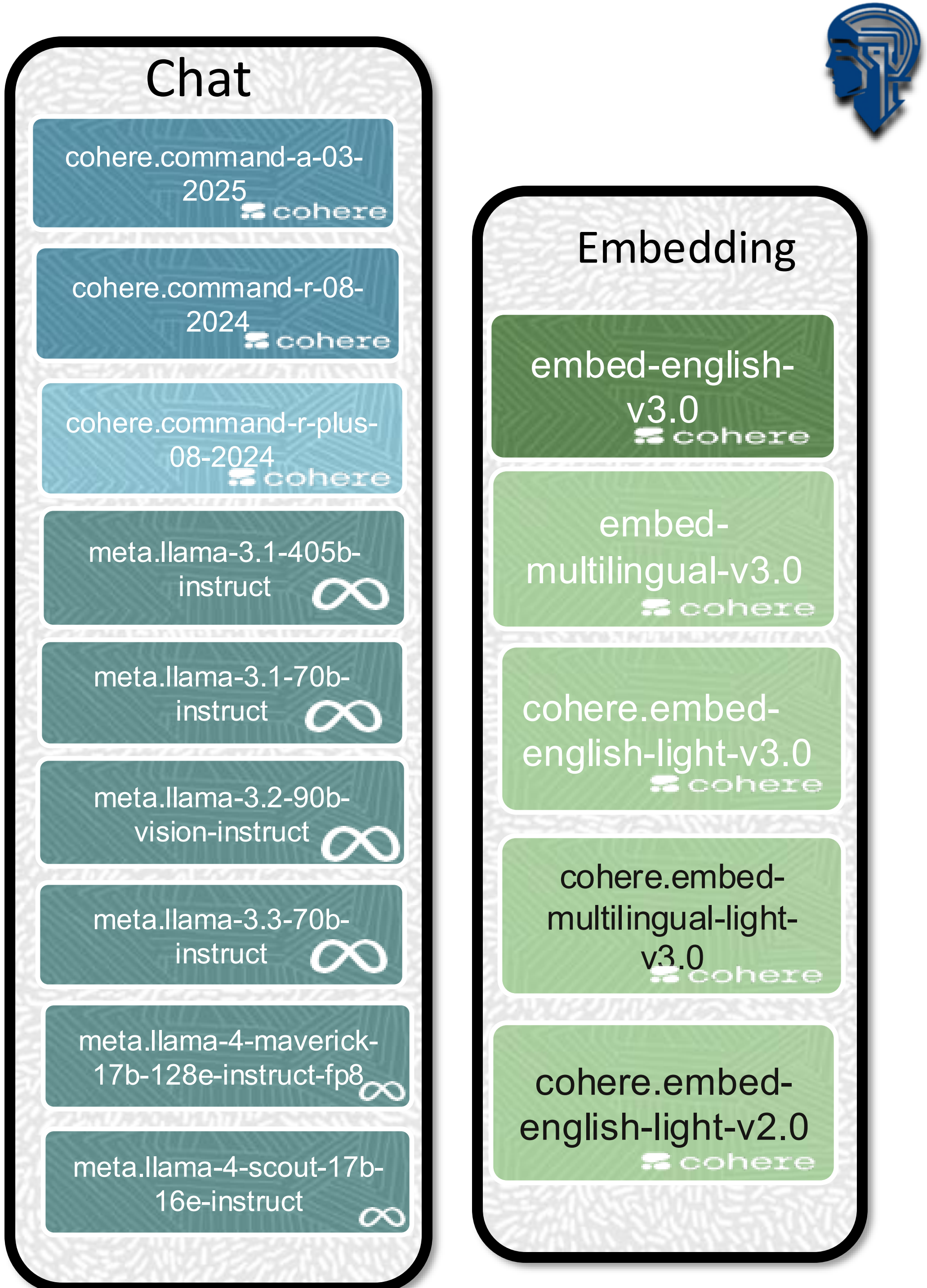
Chat Models

- Ask questions in natural language or submit text and get answers and continue with follow-up questions.
    - **Brazil East (Sao Paulo),**
    - **Germany Central (Frankfurt),**
    - **UK South (London),**
    - **US Midwest (Chicago) and**
    - **Japan Central (Osaka)**
    - **… Check for Other Regions for availability**

Embedding Models

- Convert text to vector embeddings
- Semantic Search
- Multilingual Models

xai.grok-3
xai.grok-3-fast
xai.grok-3-mini
xai.grok-3-mini-fast



Chat

- cohere.command-a-03-2025
- cohere.command-r-08-2024
- cohere.command-r-plus-08-2024
- meta.llama-3.1-405b-instruct
- meta.llama-3.1-70b-instruct
- meta.llama-3.2-90b-vision-instruct
- meta.llama-3.3-70b-instruct
- meta.llama-4-maverick-17b-128e-instruct-fp8
- meta.llama-4-scout-17b-16e-instruct

Embedding

- embed-english-v3.0
- embed-multilingual-v3.0
- cohere.embed-english-light-v3.0
- cohere.embed-multilingual-light-v3.0
- cohere.embed-english-light-v2.0

Pre-Trained Chat Models in Generative AI

- The cohere.command-r-08-2024 model,
  - Input token limit is **128,000** and output limit is 4,000.
  - Multilingual support of 10 languages: Arabic, Chinese (Mandarin), English, French, German, Italian, Japanese, Korean, Portuguese, and Spanish
- For the Meta Llama models,
  - The context length for input plus output is 128,000 tokens.

cohere.command-r-08-2024

cohere.command-r-plus-08-2024

cohere.command-a-03-2025

meta.llama-4-scout-17b-16e-instruct

meta.llama-3.1-70b-instruct

meta.llama-3.1-405b-instruct

meta.llama-3.2-90b-vision-instruct

meta.llama-3.3-70b-instruct

Meta Llama 3.1

- The meta.llama-3.1-405b-instruct and meta.llama-3.1-70b-instruct
- key features:
  - Model Sizes: 405 and 70 billion parameters
  - Context Length: 128,000 tokens, which is 16 times increase from the Meta Llama 3 models
  - Multilingual Support: English, French, German, Hindi, Italian, Portuguese, Spanish, and Thai
- The meta.llama-3.1-405b-instruct
  - This is a high-performance option that offers speed and scalability.
  - Compared to the meta.llama-3.1-70b-instruct model, it can handle a higher volume of requests and support more complex use cases.

Meta Llama 3.1

The meta.llama-3.1-405b-instruct Key features :

- Suited for enterprise-level applications and R&D initiatives.
- Shows exceptional capabilities in areas such as general knowledge, synthetic data generation, advanced reasoning and long-form text, multilingual translation, coding, math, and tool use.

The meta.llama-3.1-70b-instruct

- This 70 billion-parameter generation model is perfect for content creation, conversational AI, and enterprise applications.
  - Summarizing, rewording, and classifying text with high accuracy
  - Sentiment analysis and language modeling capabilities
  - Effective dialogue systems
  - Code generation

Meta Llama 3.2

Meta Llama 3.2 90B Vision

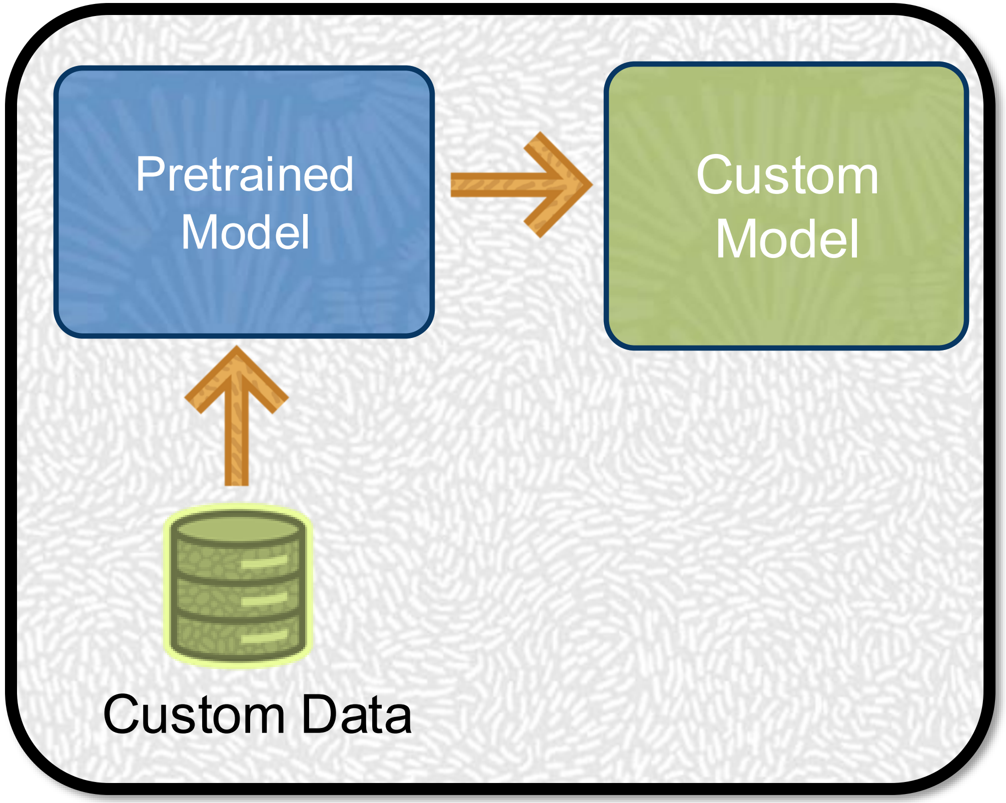
Key features

- Multimodal support (new feature): Vision support for image understanding
- Model Sizes: 90 billion parameters
- Context Length: 128,000 tokens
- Multilingual Support: English, French, German, Hindi, Italian, Portuguese, Spanish, and Thai

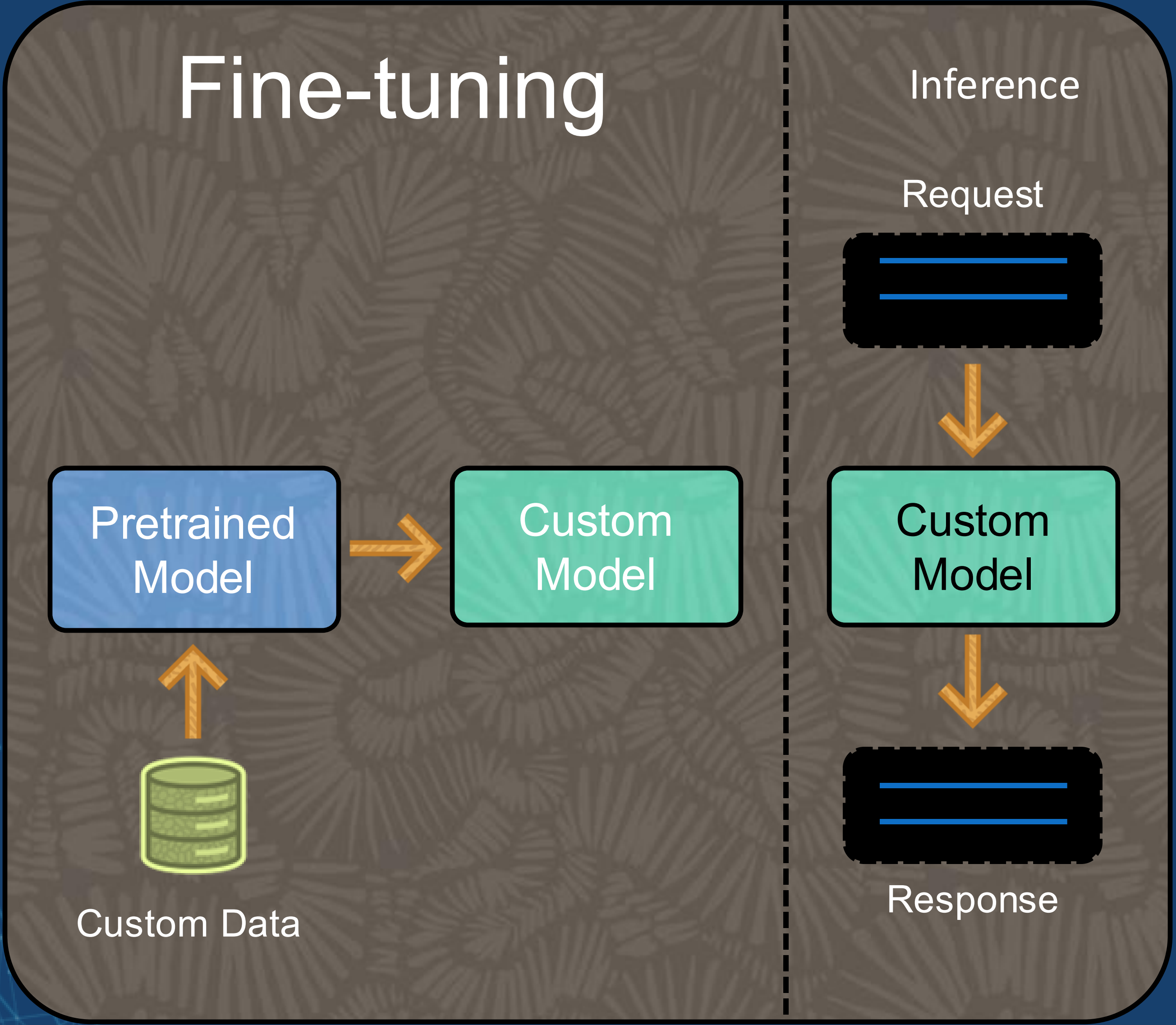
- Submit an image, ask questions about the image, and get a text outputs such as:
    - Advanced image captions
    - Detailed description of an image.
    - Answers to questions about an image.
    - Information about charts and graphs in an image.

Fine-tuning

- Optimizing a pretrained foundational model on a smaller domain-specific dataset.
  - Improve Model Performance on specific tasks
  - Improve Model Efficiency
- Use when a pretrained model doesn't perform your task well or you want to teach it something new.
- T-Few fine-tuning (Cohere) enables fast and efficient customizations.



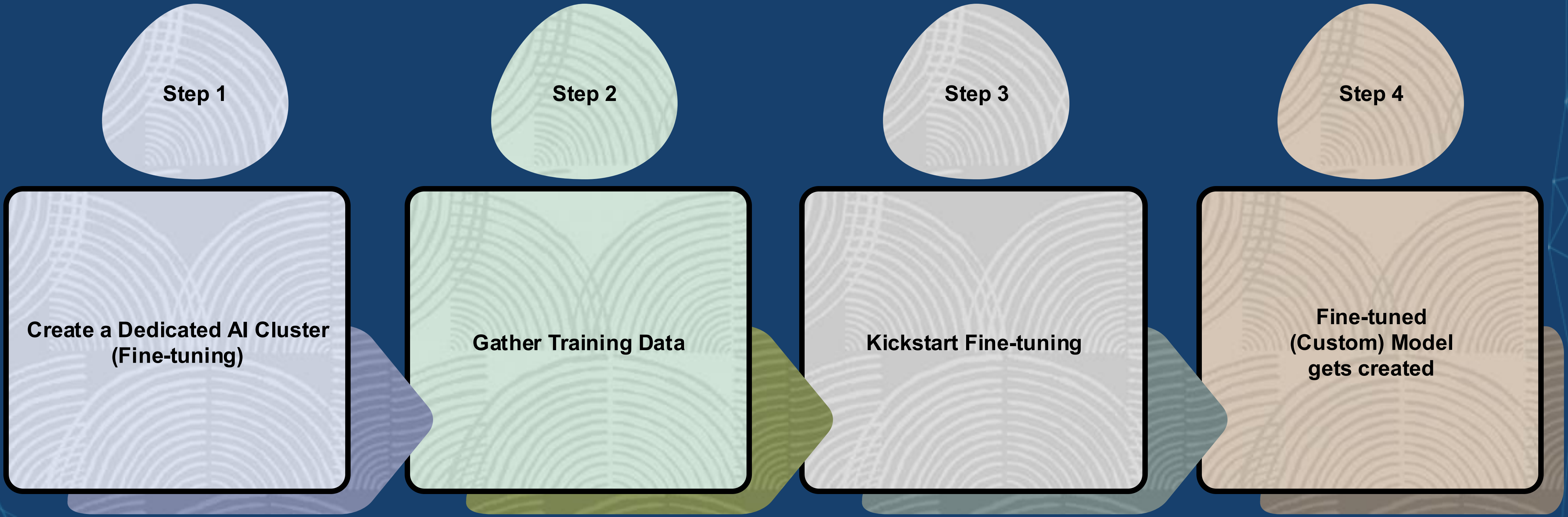
Fine-tuning and Inference



- A model is fine-tuned by taking a pretrained foundational model and providing additional training using custom data.
- In Machine Learning, Inference refers to the process of using a trained ML model to make predictions or decisions based on new input data.
- With language models, inference refers to the model receiving new text as input and generating output text based on what it has learned during training and fine-tuning.

Fine-tuning workflow in OCI Generative AI

Custom Model: A model that you can create by using a **Pretrained Model** as a base and using your own **dataset** to fine-tune that model



Step 1

Create a Dedicated AI Cluster (Fine-tuning)

Step 2

Gather Training Data

Step 3

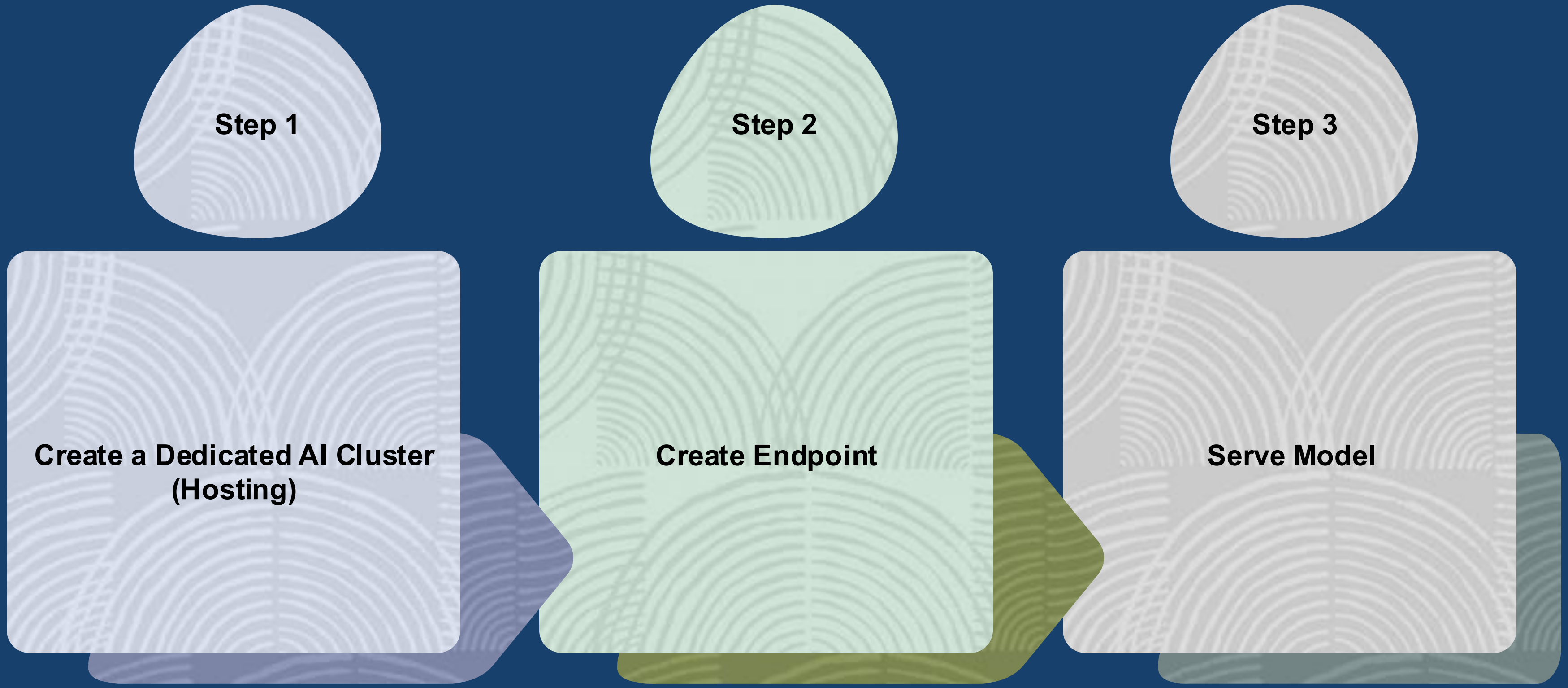
Kickstart Fine-tuning

Step 4

Fine-tuned (Custom) Model gets created

Inference workflow in OCI Generative AI

Model Endpoint: A designated point on a **Dedicated AI Cluster** where a large language model can accept user requests and send back responses such as the model's generated text



Step 1

Create a Dedicated AI Cluster (Hosting)

Step 2

Create Endpoint

Step 3

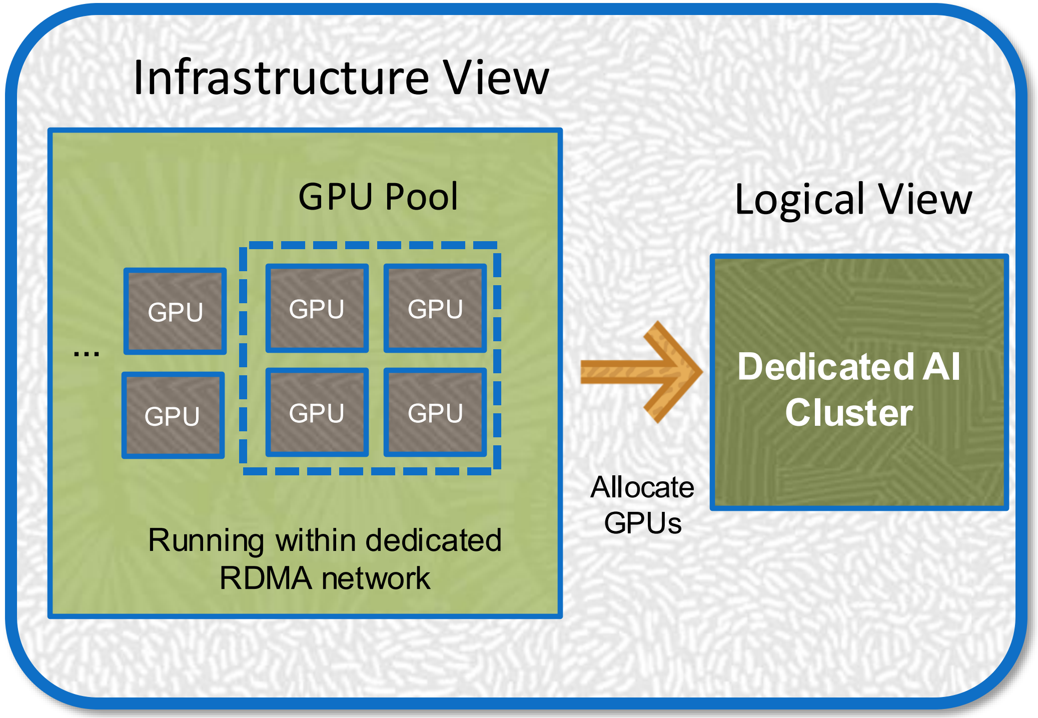
Serve Model

Dedicated AI Clusters

- Dedicated AI clusters are GPU based compute resources that host the customer's fine-tuning and inference workloads.

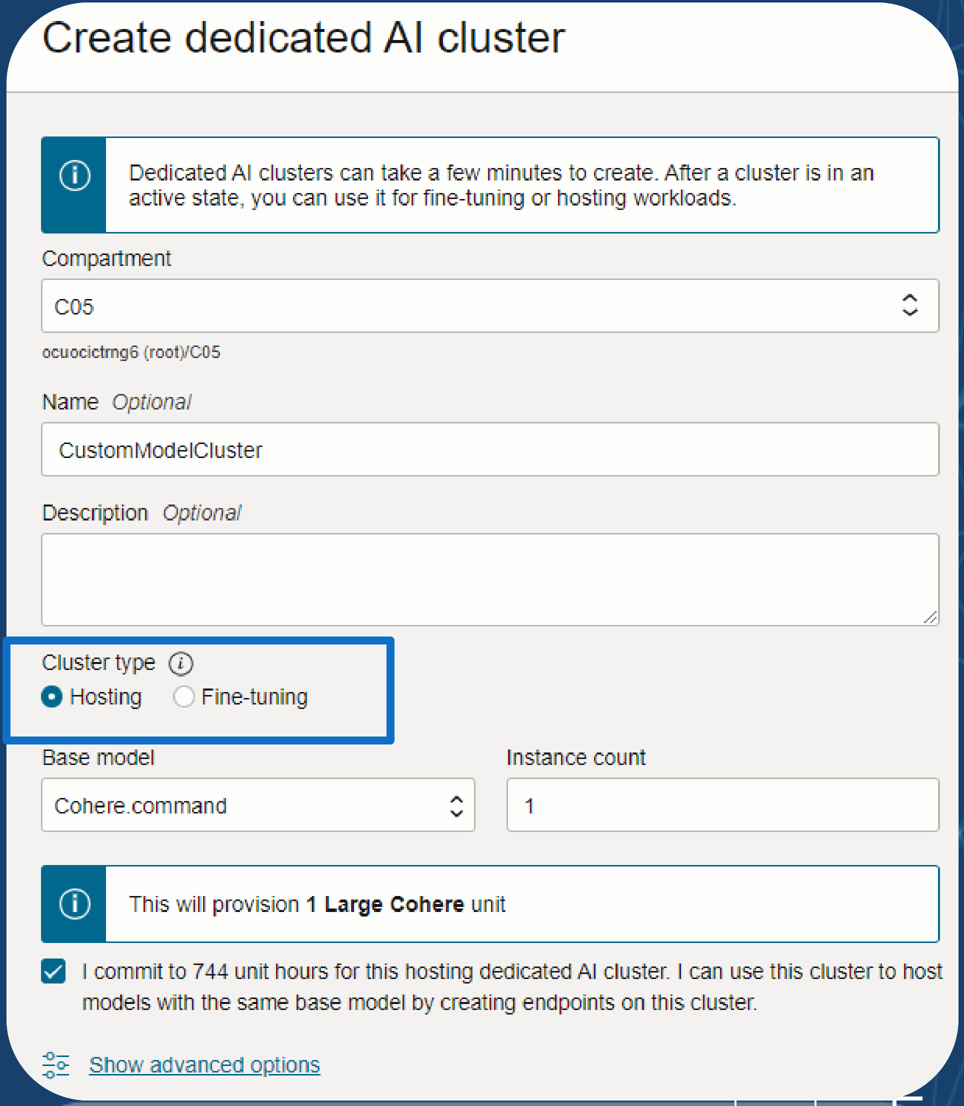
- Generative AI service establishes a dedicated AI cluster, which includes dedicated GPUs and an exclusive RDMA cluster network for connecting the GPUs.

- The GPUs allocated for a customer's generative AI tasks are isolated from other GPUs.



Dedicated AI Clusters

- Effectively a single-tenant deployment where the GPUs in the cluster only host your custom models.
- Since the model endpoint isn't shared with other customers, the model throughput is consistent.
- The minimum cluster size is easier to estimate based on the expected throughput.
- Cluster Types
  - **Fine-tuning:** used for *training* a pretrained foundational model
  - **Hosting**: used for hosting a custom model endpoint for *inference*



Create dedicated AI cluster

ⓘ Dedicated AI clusters can take a few minutes to create. After a cluster is in an active state, you can use it for fine-tuning or hosting workloads.

Compartment: C05

ocuocictrng6 (root)/C05

Name *Optional*: CustomModelCluster

Description *Optional*:

Cluster type ⓘ

● Hosting ○ Fine-tuning

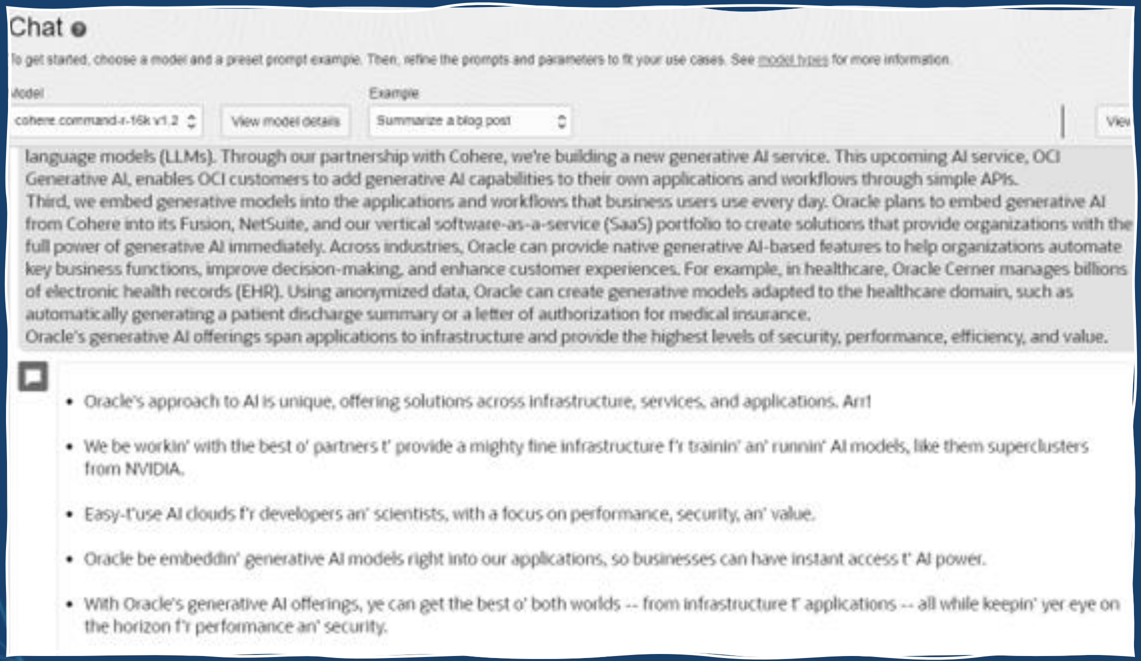
Base model: Cohere.command

Instance count: 1

ⓘ This will provision **1 Large Cohere** unit

☑ I commit to 744 unit hours for this hosting dedicated AI cluster. I can use this cluster to host models with the same base model by creating endpoints on this cluster.

Chat Model Parameters



Chat

To get started, choose a model and a preset prompt example. Then, refine the prompts and parameters to fit your use cases. See model types for more information.

Model: cohere.command-r-16k v1.2

Example: Summarize a blog post

language models (LLMs). Through our partnership with Cohere, we're building a new generative AI service. This upcoming AI service, OCI Generative AI, enables OCI customers to add generative AI capabilities to their own applications and workflows through simple APIs.

Third, we embed generative models into the applications and workflows that business users use every day. Oracle plans to embed generative AI from Cohere into its Fusion, NetSuite, and our vertical software-as-a-service (SaaS) portfolio to create solutions that provide organizations with the full power of generative AI immediately. Across industries, Oracle can provide native generative AI-based features to help organizations automate key business functions, improve decision-making, and enhance customer experiences. For example, in healthcare, Oracle Cerner manages billions of electronic health records (EHR). Using anonymized data, Oracle can create generative models adapted to the healthcare domain, such as automatically generating a patient discharge summary or a letter of authorization for medical insurance.

Oracle's generative AI offerings span applications to infrastructure and provide the highest levels of security, performance, efficiency, and value.

- Oracle's approach to AI is unique, offering solutions across infrastructure, services, and applications. Arr!
- We be workin' with the best o' partners t' provide a mighty fine infrastructure f'r trainin' an' runnin' AI models, like them superclusters from NVIDIA.
- Easy-t'use AI clouds f'r developers an' scientists, with a focus on performance, security, an' value.
- Oracle be embeddin' generative AI models right into our applications, so businesses can have instant access t' AI power.
- With Oracle's generative AI offerings, ye can get the best o' both worlds -- from infrastructure t' applications -- all while keepin' yer eye on the horizon f'r performance an' security.

Maximum Output Tokens

Max number of tokens model generates per response.

Preamble Override

An initial guideline message that can change the model's overall chat behavior and conversation style.

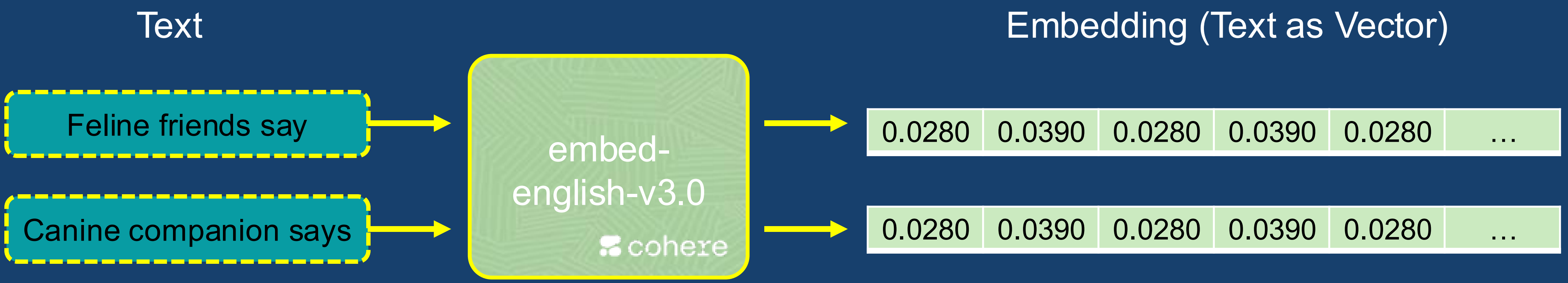
If specified, the model's default preamble is replaced with the provided preamble.

Temperature

Controls the randomness of the output. To generate the same output for a prompt every time you run that prompt, use 0.

Lower values are used in tasks with a “correct” answer (Q&A). Higher values enable the model to generate more “creative” outputs but might generate hallucinations.

Embedding Models in Generative AI



Text

Feline friends say

Canine companion says

embed-english-v3.0

cohere

Embedding (Text as Vector)

0.0280	0.0390	0.0280	0.0390	0.0280	…
0.0280	0.0390	0.0280	0.0390	0.0280	…

- Cohere.embed-english converts English text into vector embeddings.
- Cohere.embed-english-light is the smaller and faster version of embed-english.
- Cohere.embed-multilingual is the state-of-the-art multilingual embedding model that can convert text in over 100 languages into vector embeddings.

Embedding Models in Generative AI

embed-english-v3.0
embed-multilingual-v3.0

- English and Multilingual
- Model creates a 1024-dimensional vector for each embedding
- Max 512 tokens per embedding

embed-english-light-v3.0
embed-multilingual-light-v3.0

- Smaller, faster version; English and Multilingual
- Model creates a 384-dimensional vector for each embedding.
- Max 512 tokens per embedding

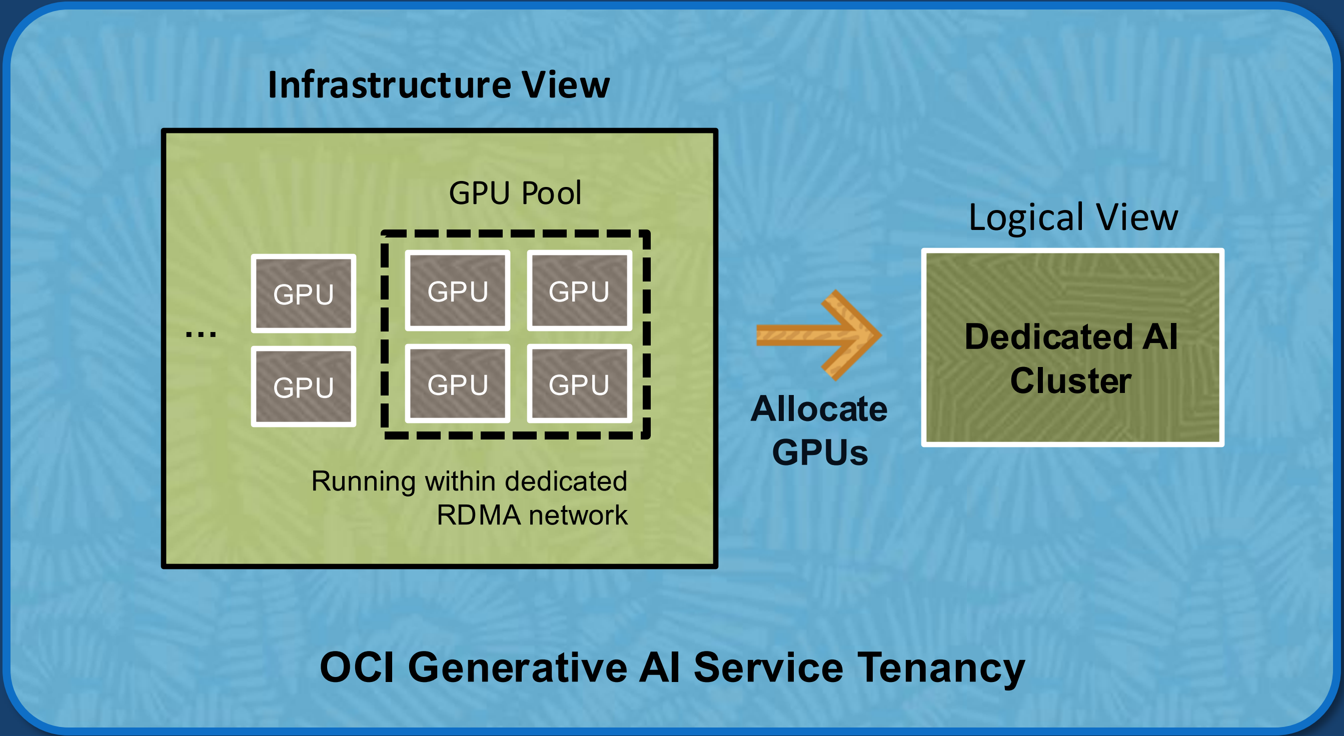
embed-english-light-v2.0

- Previous generation models, English
- Model creates a 384 dimensional vector for each embedding
- Max 512 tokens per embedding

OCI Generative AI Security

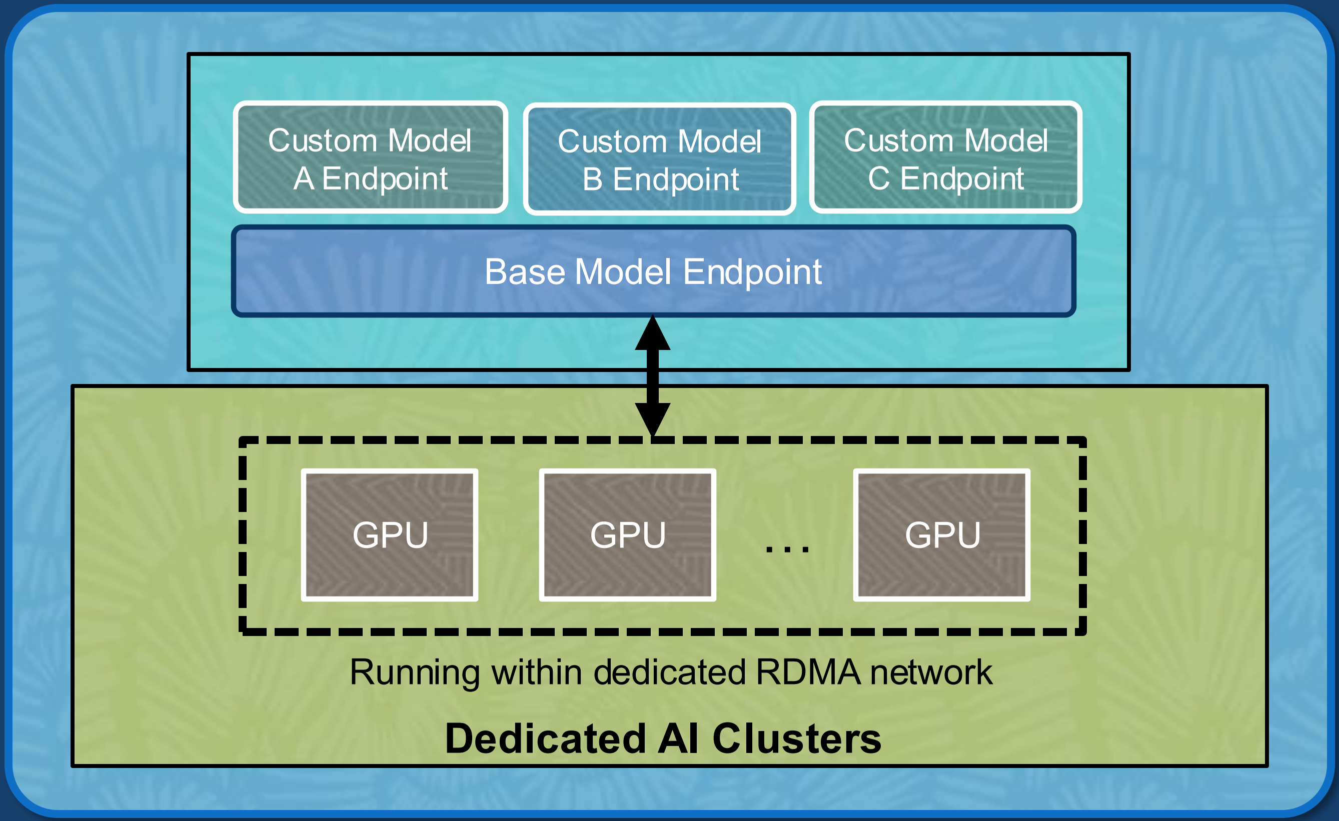
Dedicated GPU and RDMA Network

- Security and privacy of customer workloads is an essential design tenet.
- GPUs allocated for a customer's generative AI tasks are isolated from other GPUs.



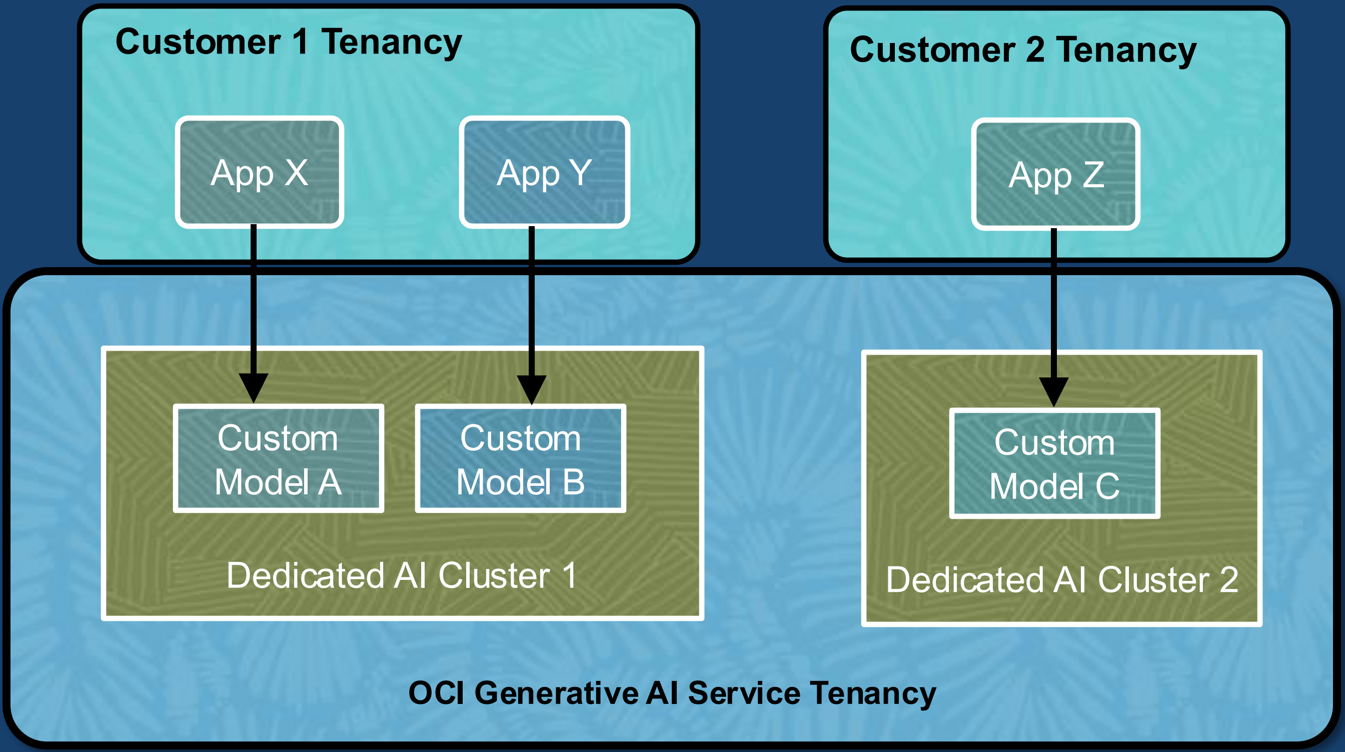
Model Endpoints

- For strong data privacy and security, a dedicated GPU cluster only handles fine-tuned models of a single customer.
- Base model + fine-tuned model endpoints share the same cluster resources for the most efficient utilization of underlying GPUs in the dedicated AI cluster.



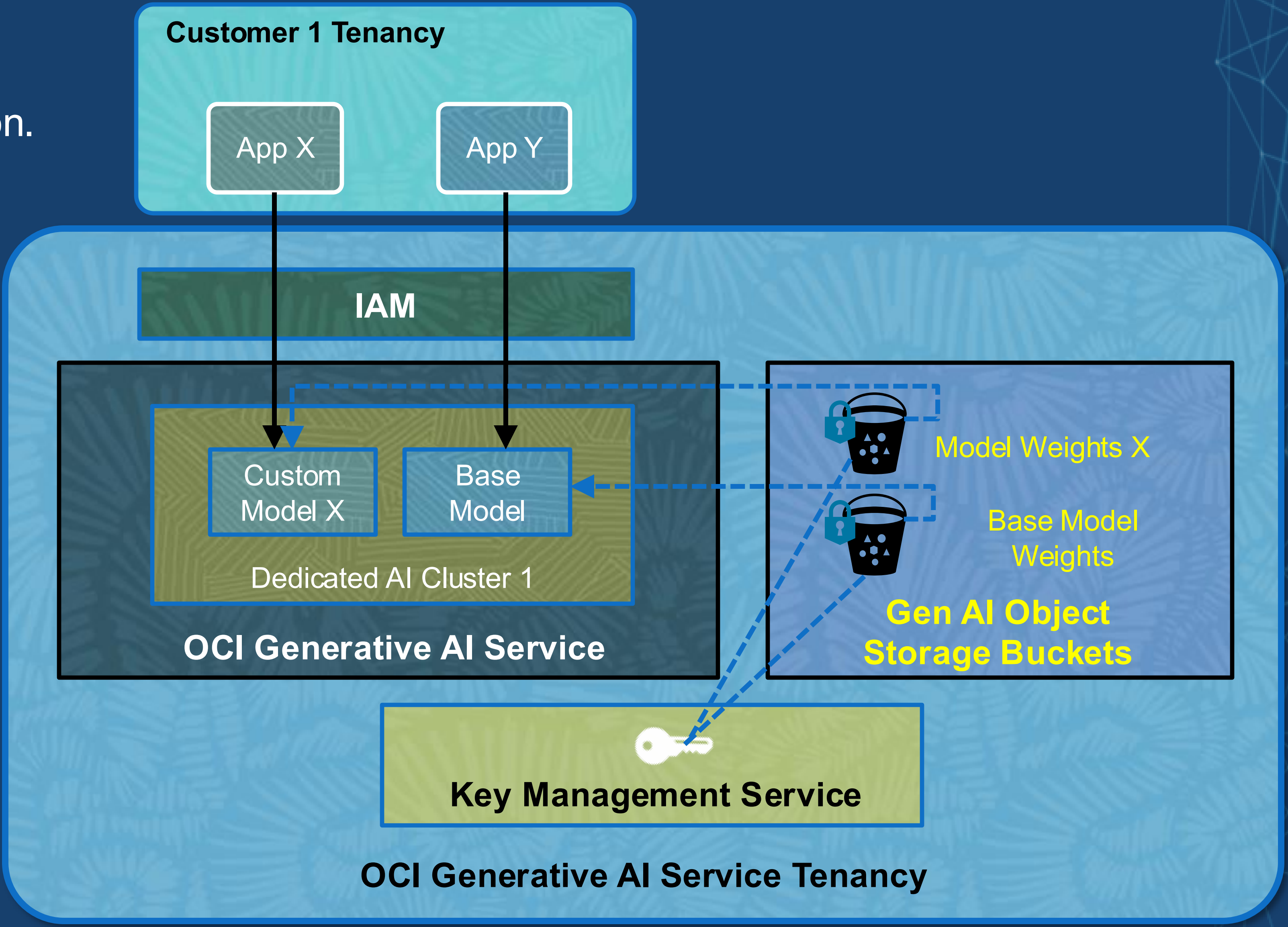
Customer Data and Model Isolation

- Customer data access is restricted within the customer’s tenancy, so that one customer’s data can’t be seen by another customer.
- Only a customer’s application can access custom models created and hosted from within that customer’s tenancy.



Generative AI leverages OCI Security Services

- Leverages OCI IAM for Authentication and Authorization.
- OCI Key Management Service stores base model keys securely.
- The fine-tuned customer models weights are stored in OCI Object Storage buckets (encrypted by default).



Thank You