

Future-Proofing Your IT Strategy with Nutanix AI Innovations

Mihkel Meerits / Nutanix / Cloud Architect
17.10.2024

Only Nutanix Makes Hybrid Multicloud Simple and Cost Effective

Any Workload, Any App

Enterprise Apps
Cloud Native Apps
Analytics / ML
Databases
Desktops



Run Anywhere

Data Centers
Public Clouds
Service Provider Clouds
Edge Locations

One Platform to Run Apps and Data Anywhere

What is Nutanix Cloud Platform?

Nutanix Cloud Platform

Enterprise Apps

Modern Apps

Analytics AI/ML

Databases

Desktops

Unified Control Plane

Federated Management

APIs | LCM | IAM

Cloud Management

Intelligent Operations | Self-Service | Cost | Security Operations

Data Governance

Security | Privacy | Compliance

Files, Objects

Data Services for
Kubernetes

Database Service

Platform Services
Project Beacon

Cloud Infrastructure

Virtual Compute, Storage, Networking; Disaster Recovery; Security

AI-Enabled Edge

Private Cloud

MSPs

Extension to Public
Cloud

Public Clouds
(Native)

Application and data portability



Enterprise AI

What **specific business challenge** are you looking to solve with GenAI?

**What is
Enterprise AI**
and how is it
different?



It's GenAI
**focused on
business
outcomes.**

Enterprise AI for
Business? Does it
Matter?

Leaders say Yes.

49%

of CEO respondents have a workforce
productivity plan that includes GenAI

Gartner.

What are you doing for enterprise AI?

We recommend these starting points.

Create Better Security

Leverage AI models for fraud detection, threat detection, alert enrichment, and automatic policy creation.



Accelerate Code and Content Creation

Enable code co-pilots, intelligent document processing, and fine-tuning models on domain-specific datasets to accelerate code and content generation.

Supercharge the customer experience

Dive into the analysis of customer feedback, provide personalized chatbots, and create a tailored engagement.

Nutanix Enables Enterprise AI with GPT-in-a-Box.

Nutanix makes Enterprise AI Easy

A Secure and Resilient
Platform for
All Your AI Data

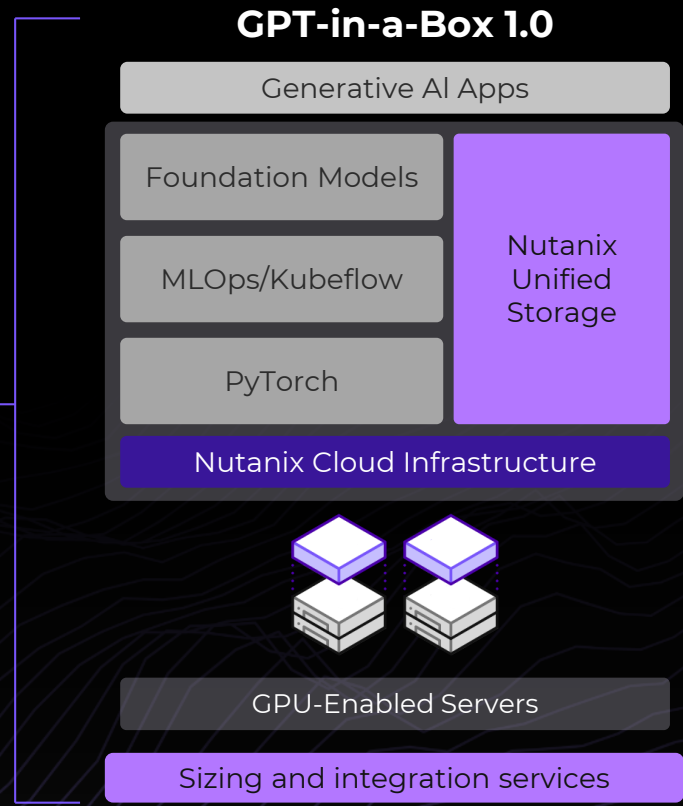
Simplify
Enterprise AI
Adoption

Reduce Your
Total Cost of
Ownership

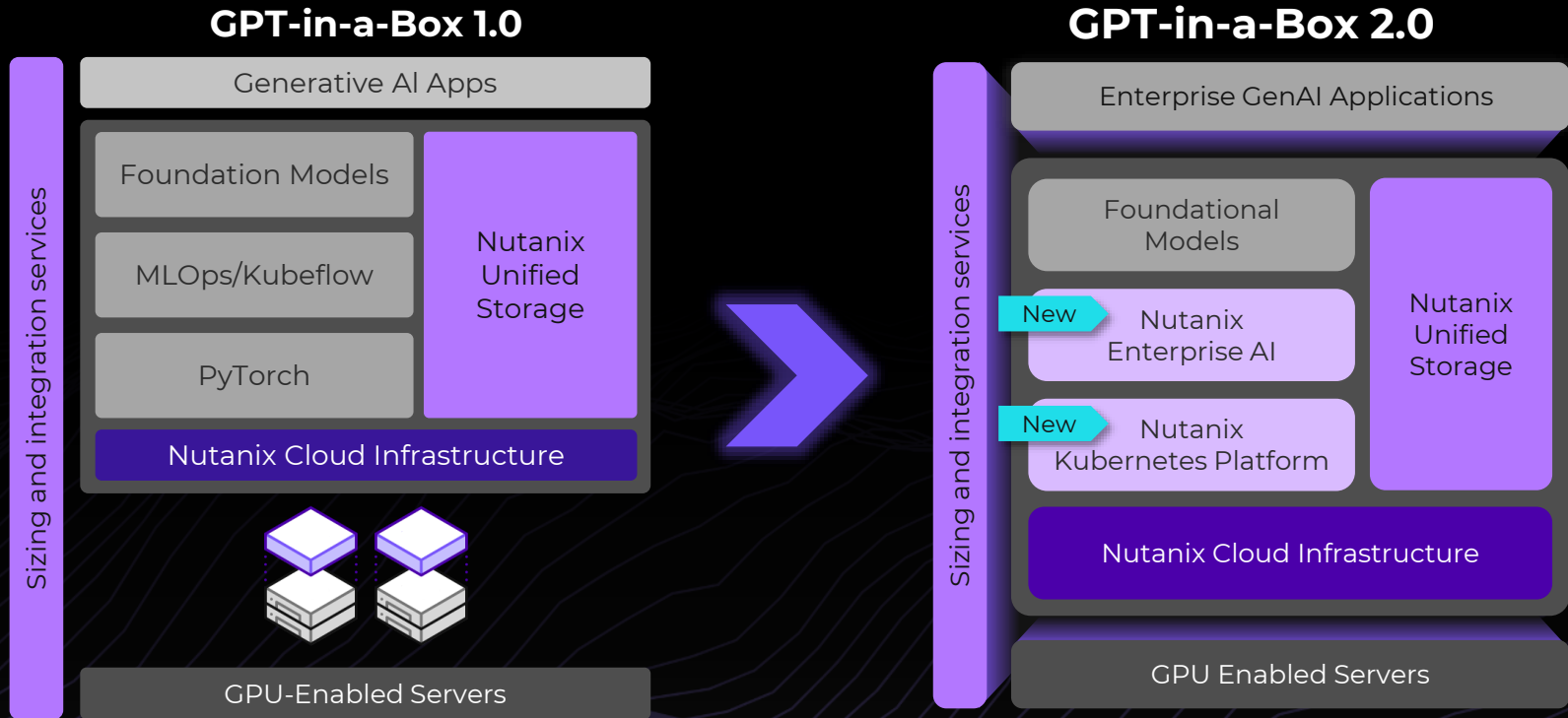


Nutanix
GPT-in-a-Box
2.0

Nutanix GPT-in-a-Box 1.0 accelerated enterprise AI adoption with an opinionated stack of open-source tools and models to get started with deploying enterprise AI.



Nutanix GPT-in-a-Box 1.0 Evolves into 2.0





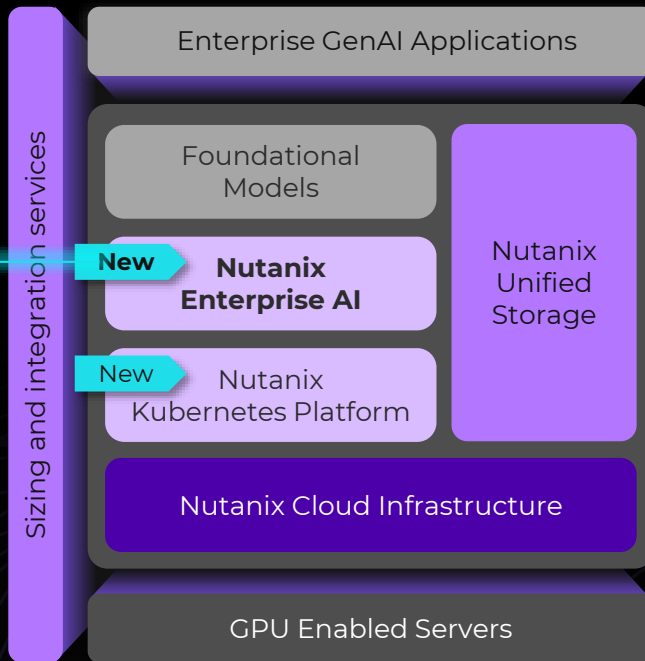
Say Hello to
Nutanix Enterprise AI



Nutanix Enterprise AI

Deploy and Operate with your choice of LLMs with secure APIs for GenAI endpoints within a simple-to-use interface you can run on any Kubernetes.

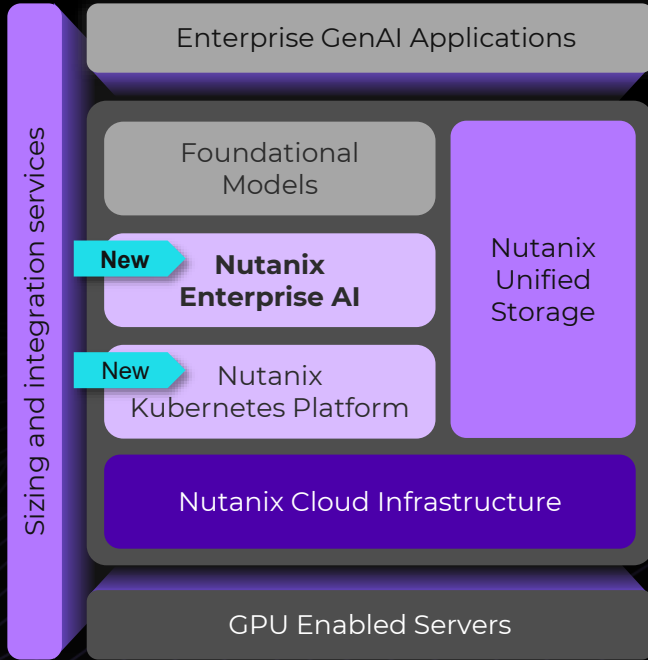
GPT-in-a-Box 2.0





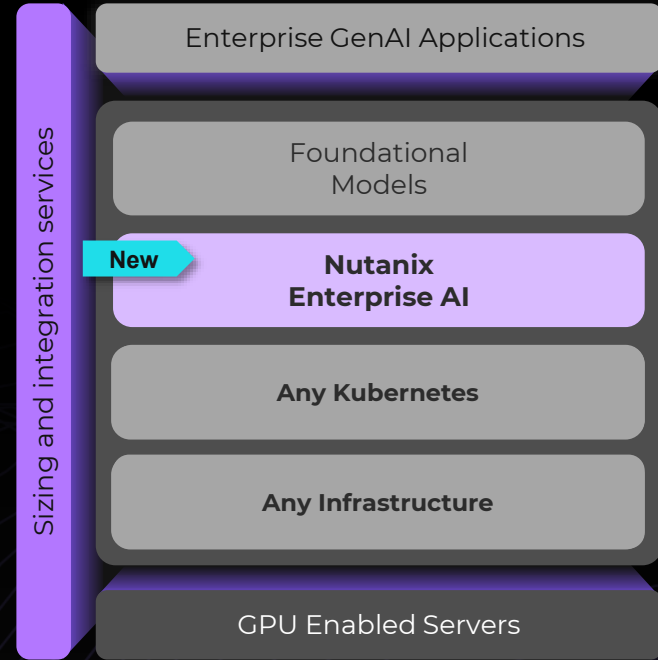
GPT-in-a-Box or Nutanix Enterprise AI?

GPT-in-a-Box 2.0



Validated Enterprise AI Stack

Nutanix Enterprise AI Stack



Choose Your Own Enterprise AI Stack



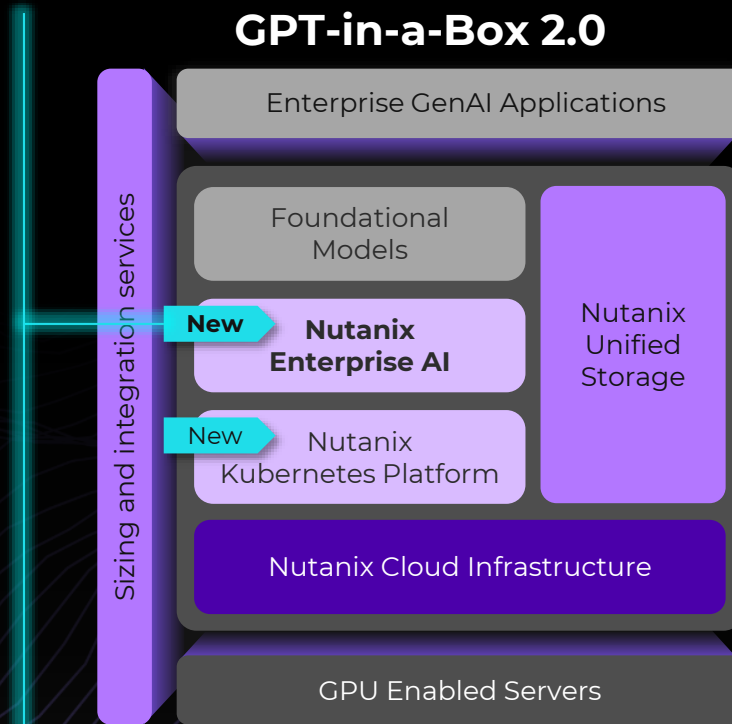
How is Nutanix Enterprise AI different?





Key Features

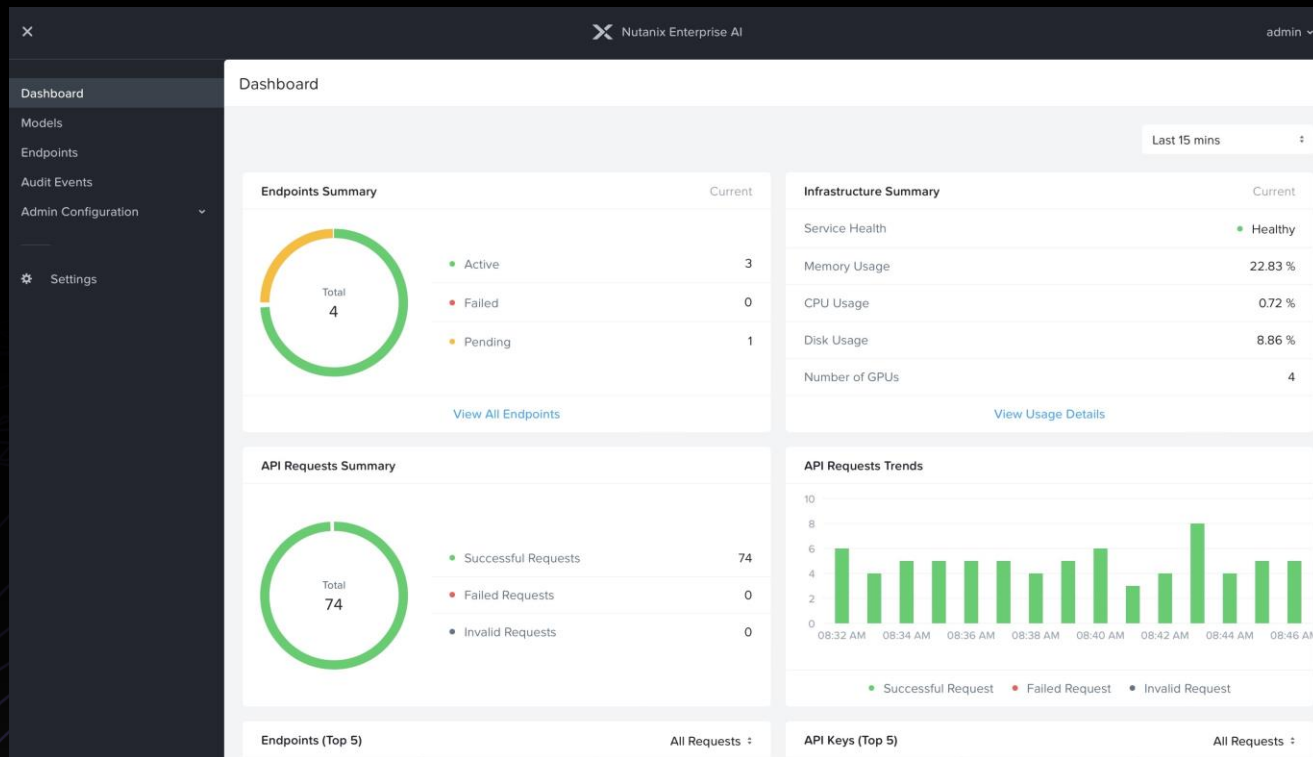
- A simple, enterprise-grade user interface
- LLM (AI model) selection and deployment from Hugging Face and NVIDIA NIM
- LLM pre-deployment confirmation testing
- LLM endpoint access control with API tokens
- API token creation/management
- API code samples
- Role-based access controls (RBAC)
- Dashboards and Monitoring
- Infrastructure health
- Event auditing
- Nutanix Pulse remote reporting





Nutanix Enterprise AI Auditing + Monitoring

Easily understand and audit infrastructure health, GPU utilization of GenAI models, users, and the impact of AI endpoints in-use





Nutanix Enterprise AI Choice of AI Models

Choose validated AI models (LLMs) from Hugging Face or NVIDIA NIM (NGC) or import a custom model of your choice.

Nutanix Enterprise AI Mike Barmonde

Models

No Models Available

Get started by importing from Hugging Face, NVIDIA NGC catalog or upload model manually. Pre-validated models can be imported from Hugging Face, NVIDIA NGC or models can be manually uploaded from a file share or a bucket.

From NVIDIA NGC Catalog
Import models which are pre-validated to run on Nutanix directly from NVIDIA NGC catalog.
[Import from NVIDIA NGC Catalog](#)

From Hugging Face Model Hub
Import models which are pre-validated to run on Nutanix directly from Hugging Face.
[Import from Hugging Face Model Hub](#)

Import Model
Package the models in a compatible format and manually upload the models from a file share or bucket.
[Upload Manually](#)



Nutanix Enterprise AI with Secure APIs

Create role-based access controls (RBAC) for secure API management tied to individual users.

Create an Endpoint

- List of active models imported by you
- Use GPUs for running the model
- Number of GPUs (Per Instance): 1
- GPU Card

API Key Details

- The key has been generated. Copy the key and store it securely. You cannot view the key again after you close this dialog box. If you lose the key, you must generate a new one.
- Key Name: mb-api-ngc-llama-31
- API Key: ea4e4c0b-ca02-4434-b0e9-811662e8950e
-
-

The endpoints can be accessed using an API Key. New Keys can be generated or existing keys can be used.

API Keys

- mb-api-ngc-llama-31

List of active API Keys created by you





Nutanix Enterprise AI Model Testing

Test AI model access with an API endpoint using a natural language interface before sending API access out to AI developers or application owners.

The screenshot shows the Nutanix Enterprise AI interface. A 'Test Endpoint' dialog box is open for the 'nim-llama3-1' model. The dialog displays the following information:

- Endpoint Name:** nim-llama3-1
- Model Instance:** NIM-llama-3.1-8b-instruct
- Requests:**
 - Sample Request
 - Custom Request
- Input Text 1:** Explain how Deep Neural Networks work in simple terms
- Input Text 2:** Describe the landscapes of Mars in a few sentences

The **Test** button is highlighted. The **Status** is **Succeeded**. The **Result** is:

Deep Neural Networks (DNNs) are a type of machine learning model that can learn and represent complex data patterns on their own. They're composed of multiple layers of "neurons" (mathematical functions) that process and transform input data.

Here's a simplified breakdown:

- Input Layer:** Your data, such as images, text, or sounds, is fed into the network.
- Hidden Layers:** These layers process the input data by applying weights (important values) and activation functions (simplifying formulas). Each layer learns to represent the data in a more abstract way, like recognizing patterns or features.
- Output Layer:** The

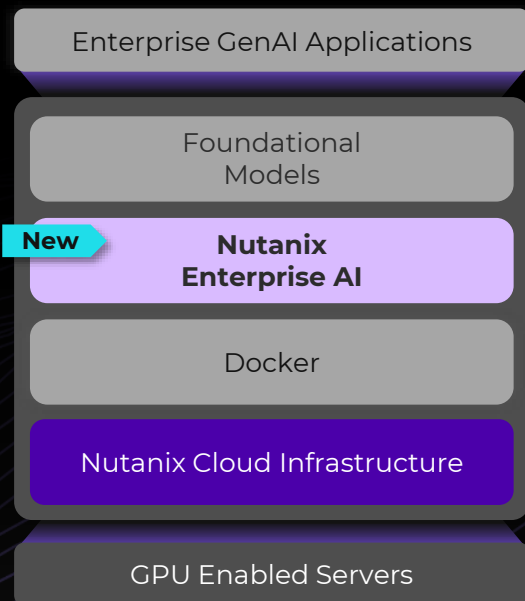
The dialog also shows a **Done** button at the bottom right. The background interface shows a sidebar with 'Dashboard', 'Models', 'Endpoints', 'Audit Events', 'Admin Configuration', and 'Settings'. The main area shows details for the 'nim-llama3-1' endpoint, including a 'Test' button and a table of requests.



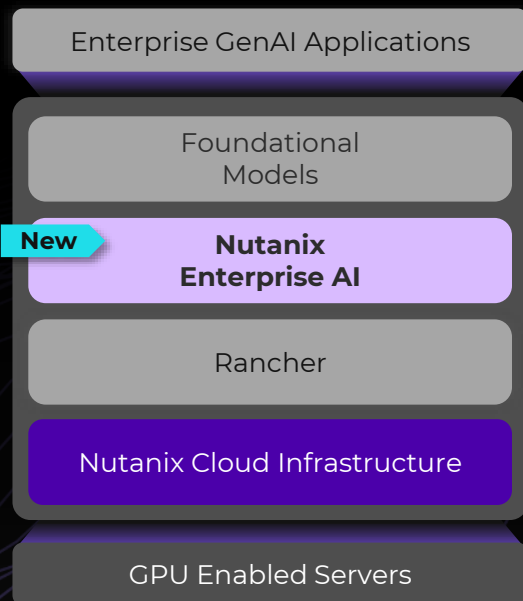
Deploy with Any Kubernetes



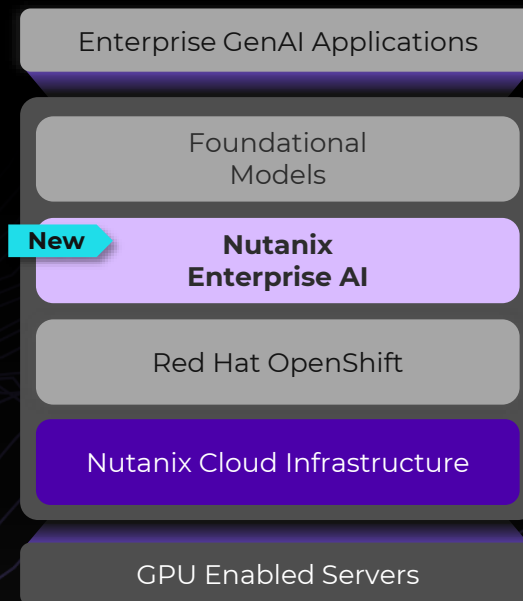
Nutanix Enterprise AI for Docker



Nutanix Enterprise AI for Rancher



Nutanix Enterprise AI for Red Hat OpenShift

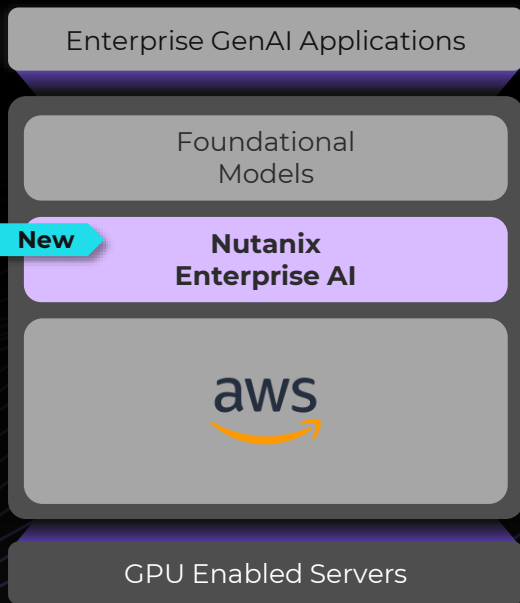




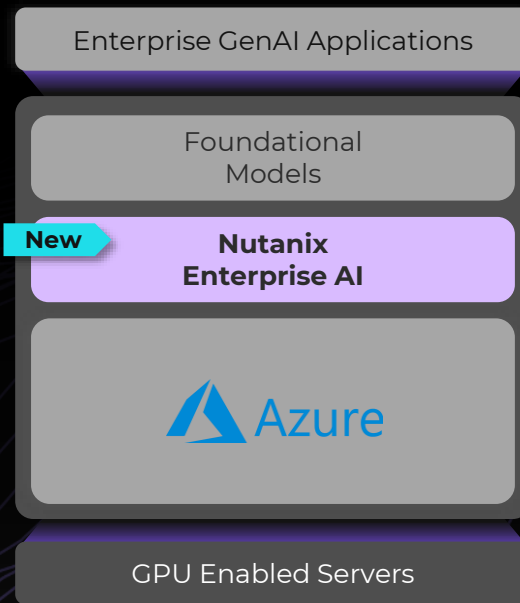
Deploy with the Leading Hyperscalers



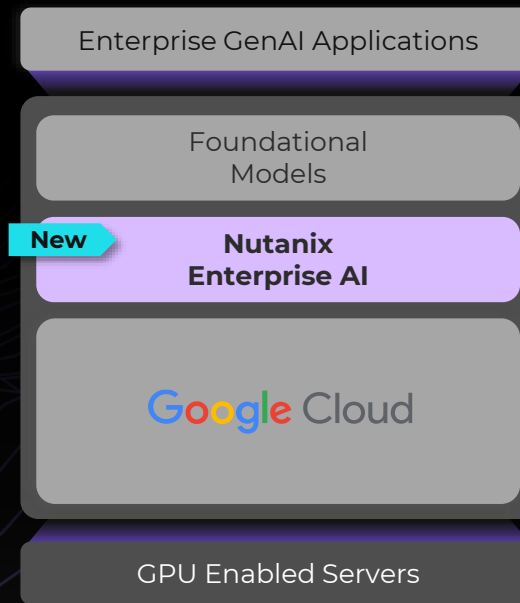
Nutanix Enterprise AI for AWS



Nutanix Enterprise AI for Microsoft Azure



Nutanix Enterprise AI for Google Cloud

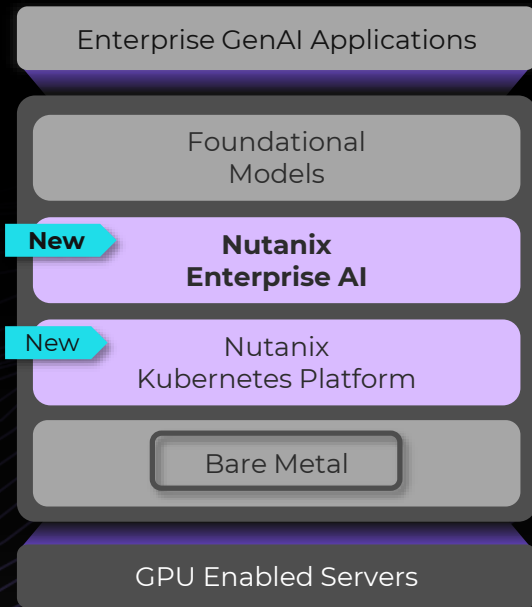




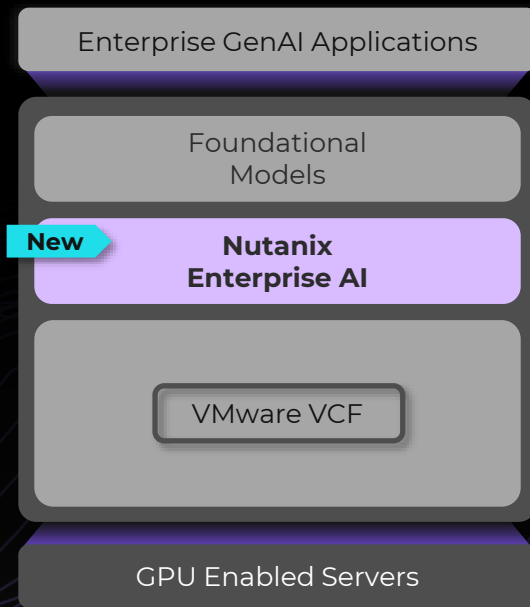
Deploy with Any Private Cloud



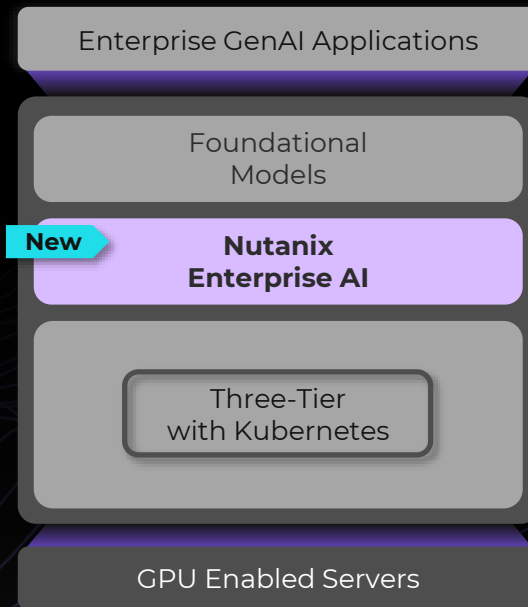
Nutanix Enterprise AI for Bare Metal



Nutanix Enterprise AI for VMware VCF



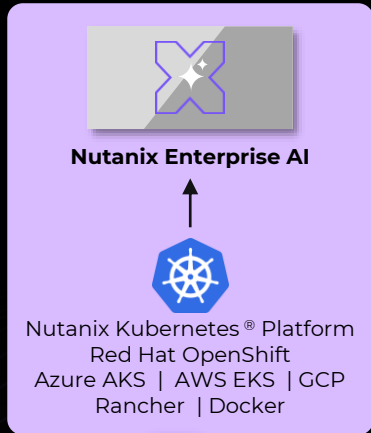
Nutanix Enterprise AI for Three-Tier





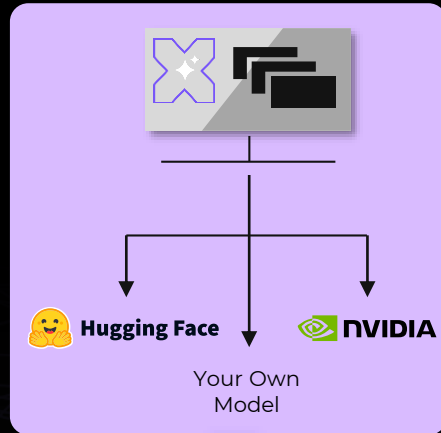
How does Nutanix Enterprise AI work?

1.



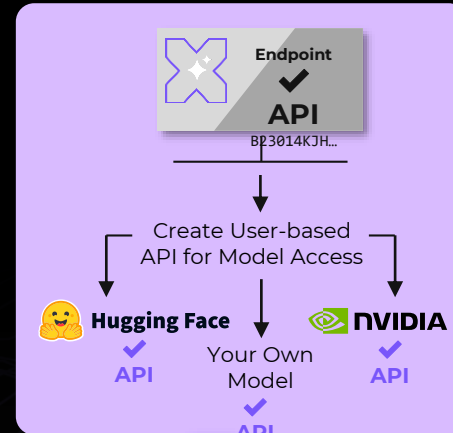
**Deploy and run
Nutanix Enterprise AI
on any Kubernetes®.**

2.



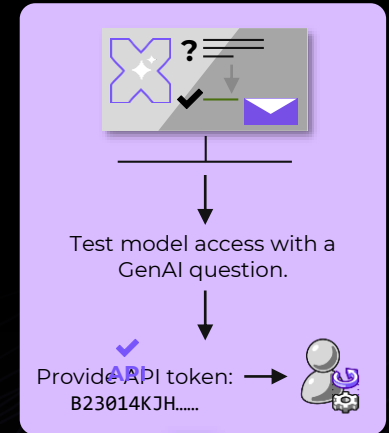
**Login to the interface
and simply pick and
deploy your LLM.**

3.



**Create a secure API to
use with your model and
test the model access.**

4.



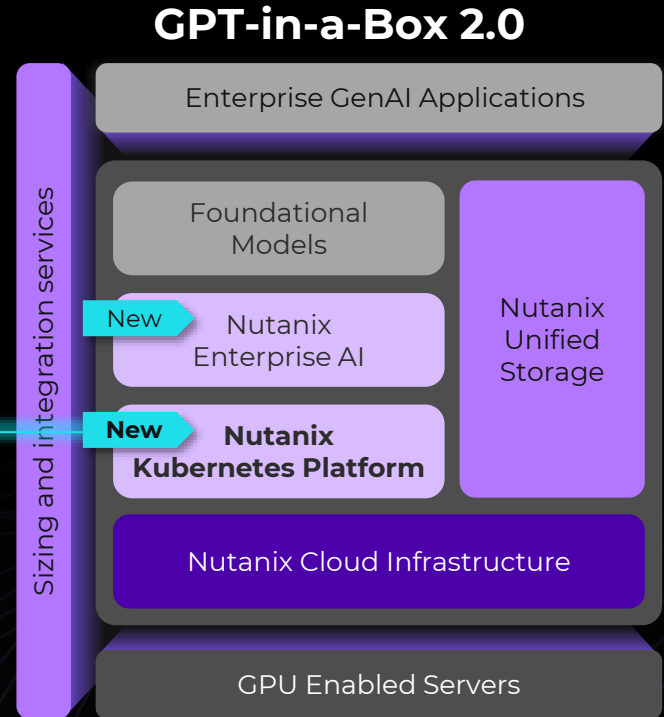
**Send the API credentials
to a developer or
application owner.**

Monitor and audit LLMs and APIs



Nutanix Kubernetes® Platform Simplifies Platform Engineering for Enterprise AI

Nutanix Kubernetes Platform accelerates AI deployment with platform engineering using a flexible, open, and secure platform complete with fleet management.



NUTANIX

Aitäh!

Mihkel Meerits / Nutanix / Cloud Architect
Mihkel.meerits@nutanix.com