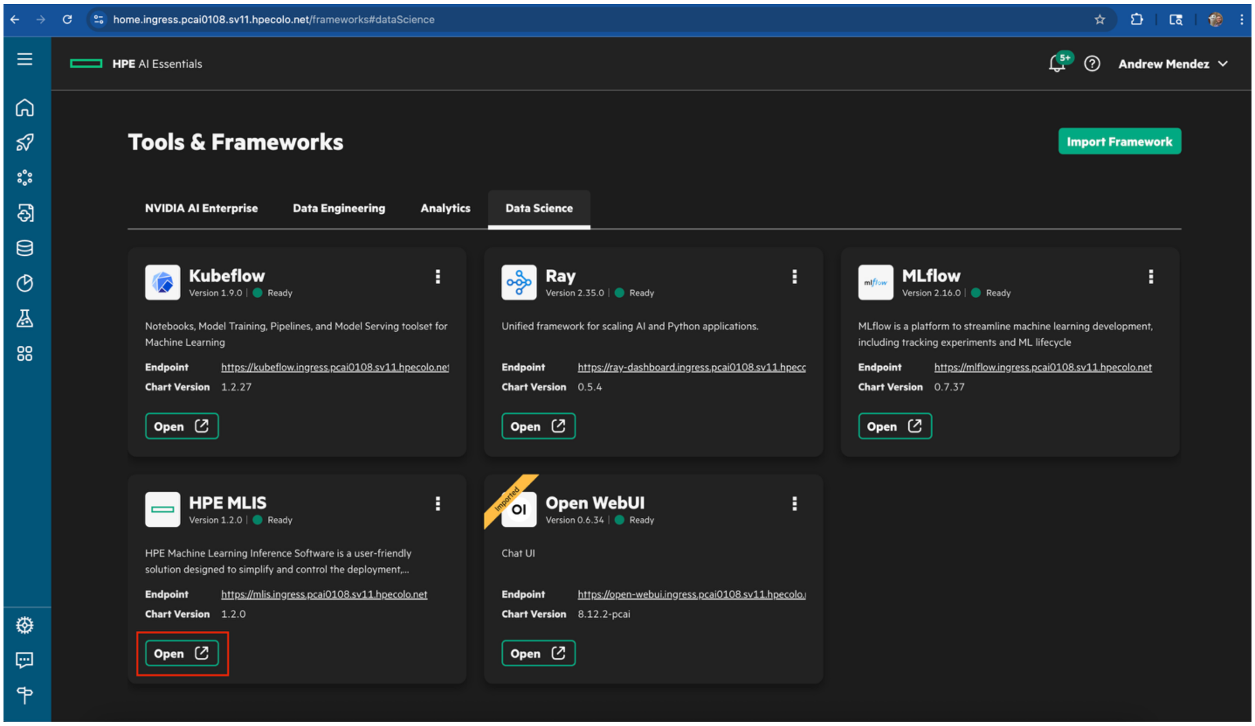
Step by Step install Helm chart

Install on AIE 1.6

Part 1: Deploy Vista model on MLIS

Go to MLIS



Tools & Frameworks

Kubeflow
Version 1.9.0 | Ready
Notebooks, Model Training, Pipelines, and Model Serving toolset for Machine Learning
Endpoint https://kubeflow.ingress.pcai0108.sv11.hpecolo.net
Chart Version 1.2.27

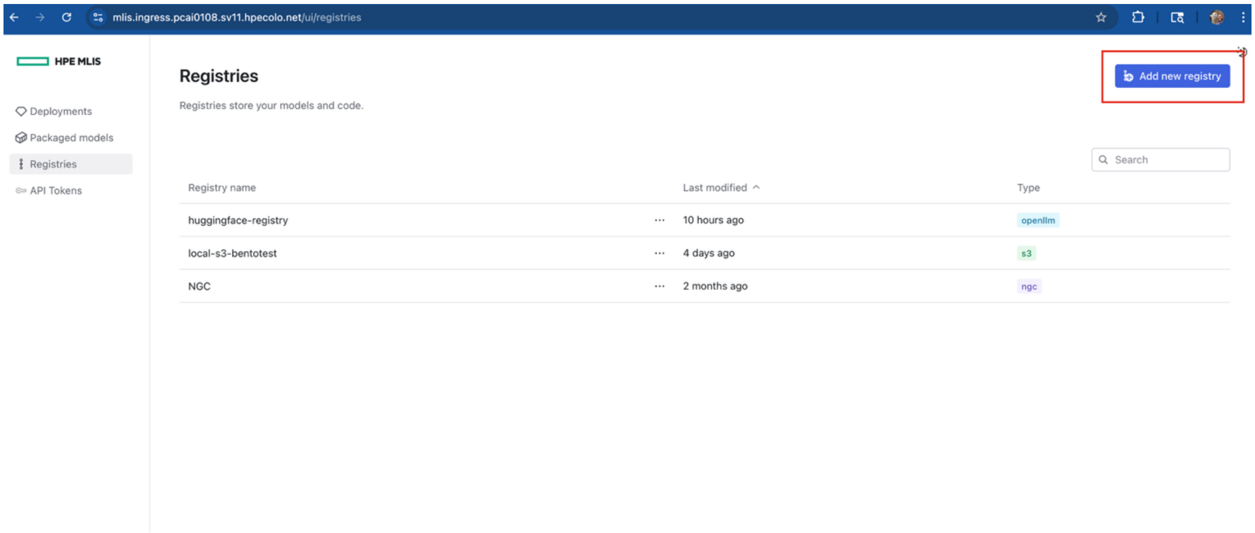
Ray
Version 2.35.0 | Ready
Unified framework for scaling AI and Python applications.
Endpoint https://ray-dashboard.ingress.pcai0108.sv11.hpecc
Chart Version 0.5.4

MLflow
Version 2.16.0 | Ready
MLflow is a platform to streamline machine learning development, including tracking experiments and ML lifecycle
Endpoint https://mlflow.ingress.pcai0108.sv11.hpecolo.net
Chart Version 0.7.37

HPE MLIS
Version 1.2.0 | Ready
HPE Machine Learning Inference Software is a user-friendly solution designed to simplify and control the deployment,...
Endpoint https://mlis.ingress.pcai0108.sv11.hpecolo.net
Chart Version 1.2.0

Open WebUI
Version 0.6.34 | Ready
Chat UI
Endpoint https://open-webui.ingress.pcai0108.sv11.hpecolo.
Chart Version 8.12.2-pcai

Next create registry select add new registry

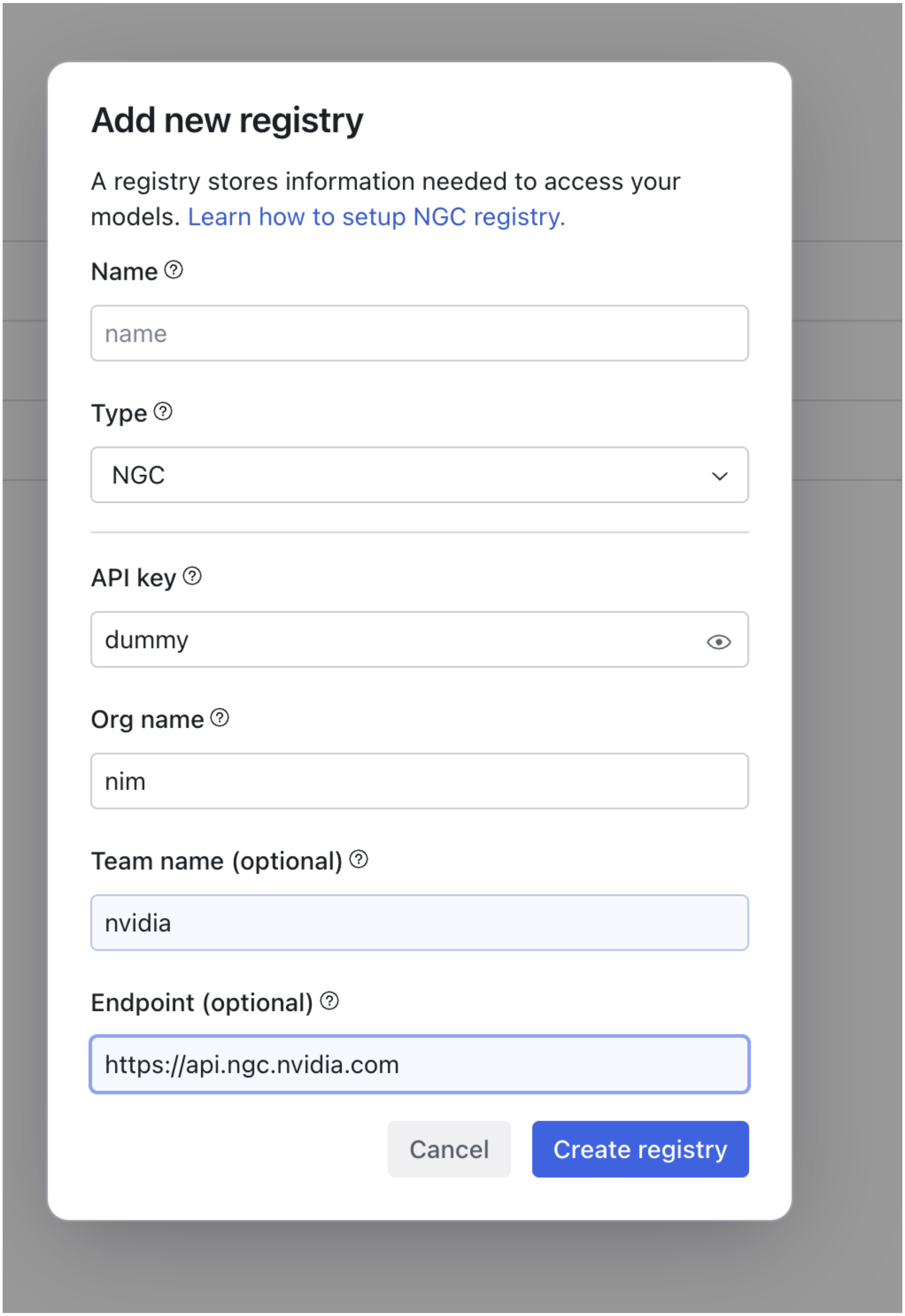


Registries

Registries store your models and code.

Registry name		Last modified ^	Type
huggingface-registry	···	10 hours ago	openllm
local-s3-bentotest	···	4 days ago	s3
NGC	···	2 months ago	ngc

Fill in the information needed to create a registry, this will hold your nvidia enterprise API key



Add new registry

A registry stores information needed to access your models. Learn how to setup NGC registry.

Name: name

Type: NGC

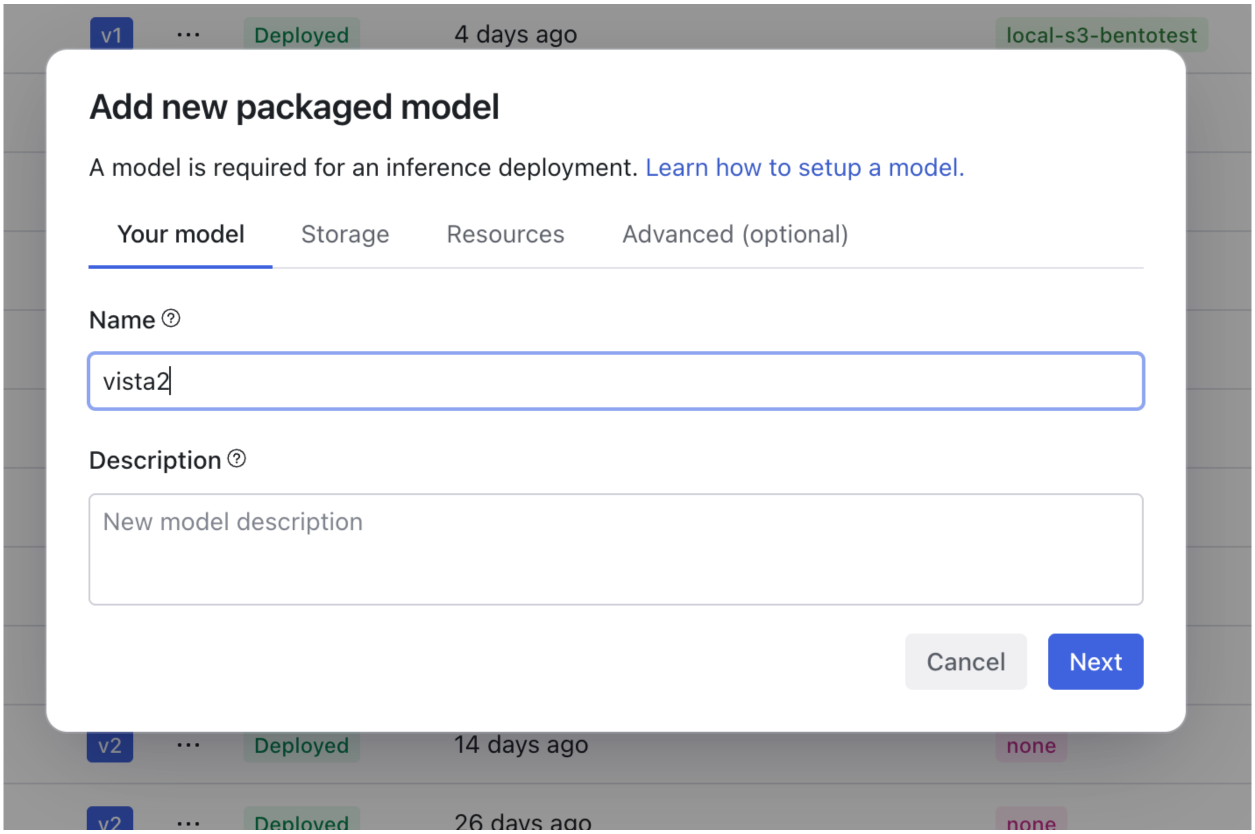
API key: dummy

Org name: nim

Team name (optional): nvidia

Endpoint (optional): https://api.ngc.nvidia.com

Next create a packaged model, name your model vista



Add new packaged model

A model is required for an inference deployment. Learn how to setup a model.

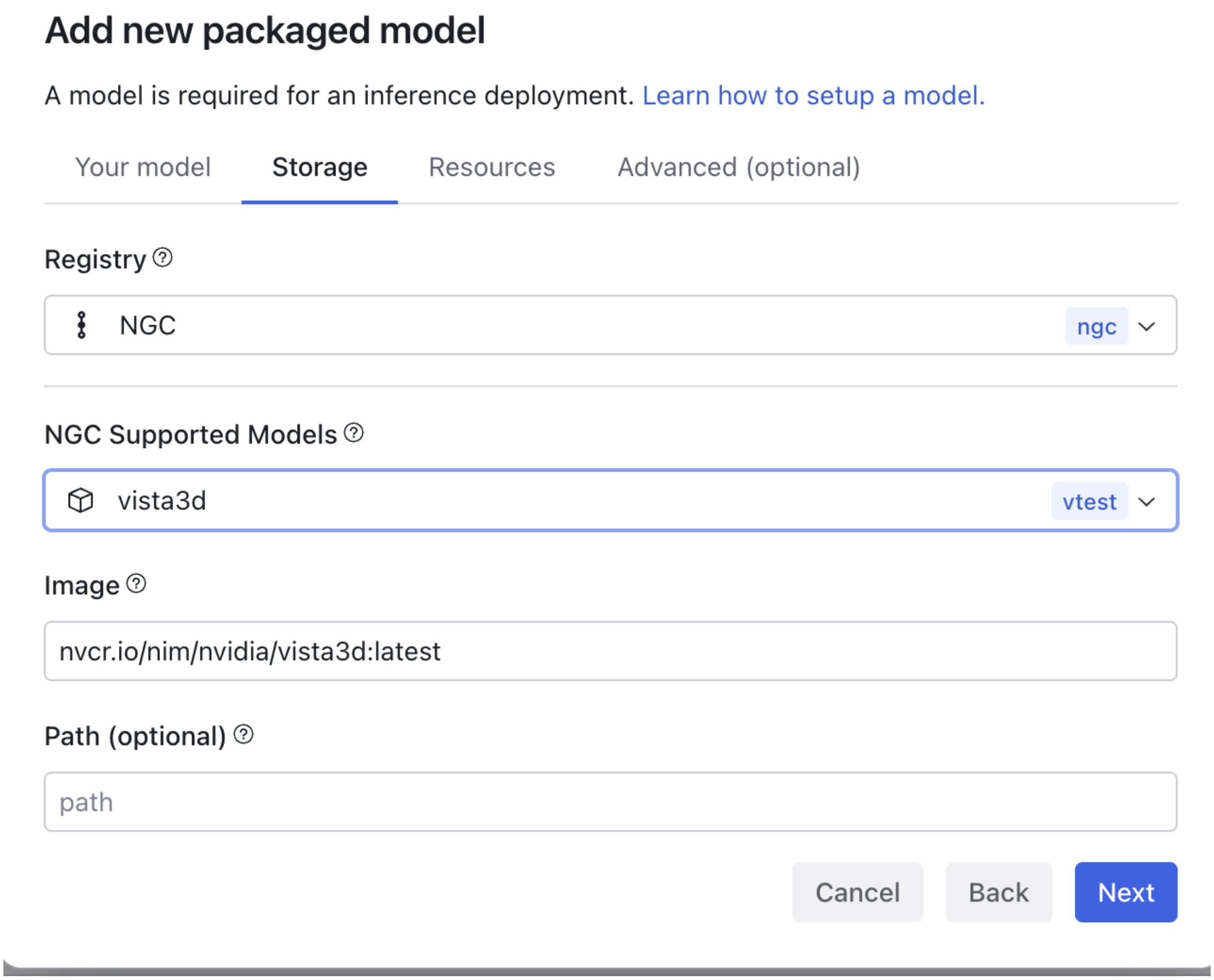
Name

vista2

Description

New model description

Select your registry and list of supported models will populate, select vista model docker image



Add new packaged model
A model is required for an inference deployment. Learn how to setup a model.
Your model
Storage
Resources
Advanced (optional)

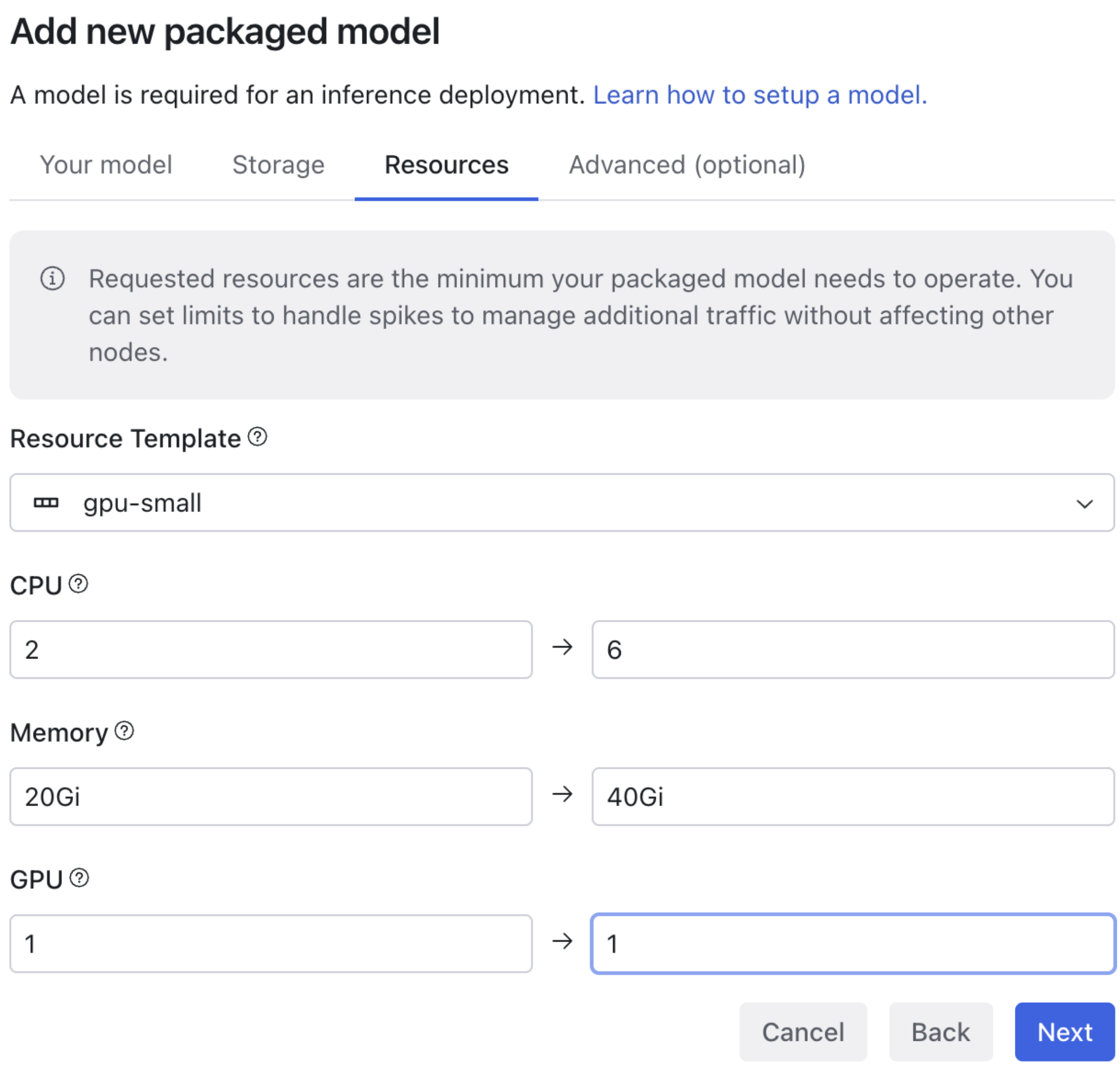
Registry: NGC (ngc)

NGC Supported Models: vista3d (vtest)

Image: nvcr.io/nim/nvidia/vista3d:latest

Path (optional): path

Next set resources to the following screenshot. Make sure to manually change gpu to 1 and 1, this model only needs 1 l40s GPU



Add new packaged model

A model is required for an inference deployment. Learn how to setup a model.

ⓘ Requested resources are the minimum your packaged model needs to operate. You can set limits to handle spikes to manage additional traffic without affecting other nodes.

Resource Template: gpu-small

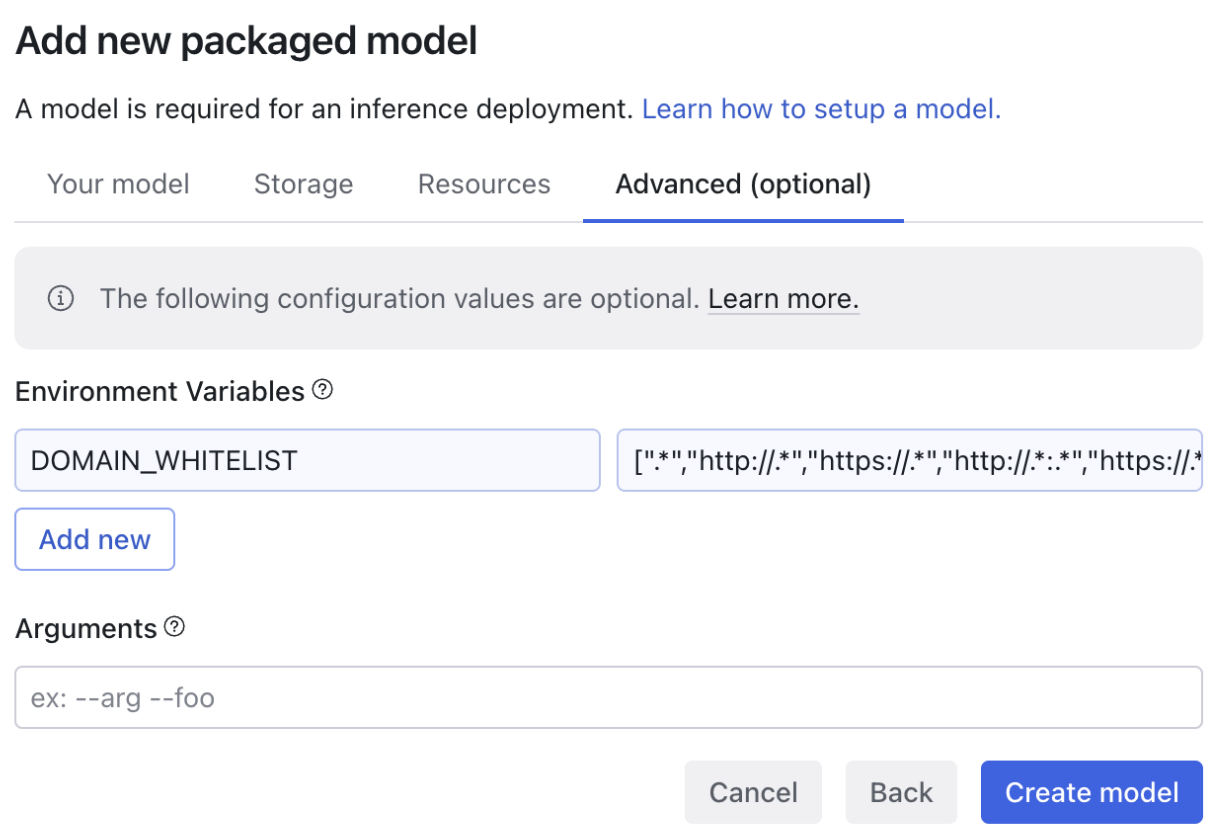
CPU: 2 → 6

Memory: 20Gi → 40Gi

GPU: 1 → 1

Next in the advanced settings, Add environment variable

DOMAIN_WHITELIST
[".*","http://.*","https://.*","http://.*:.*","https://.*:.*","file:///.*","*"]



Add new packaged model

A model is required for an inference deployment. Learn how to setup a model.

ⓘ The following configuration values are optional. Learn more.

Environment Variables

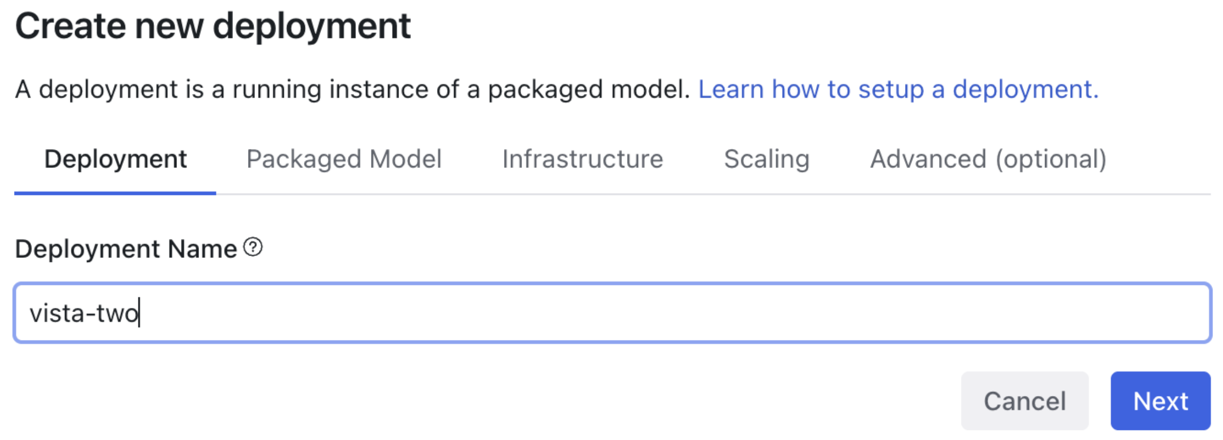
DOMAIN_WHITELIST

`[".*","http://.*","https://.*","http://.*:.*","https://.*`

Arguments

ex: --arg --foo

Now lets create a deployment, name your deployment same as packaged model



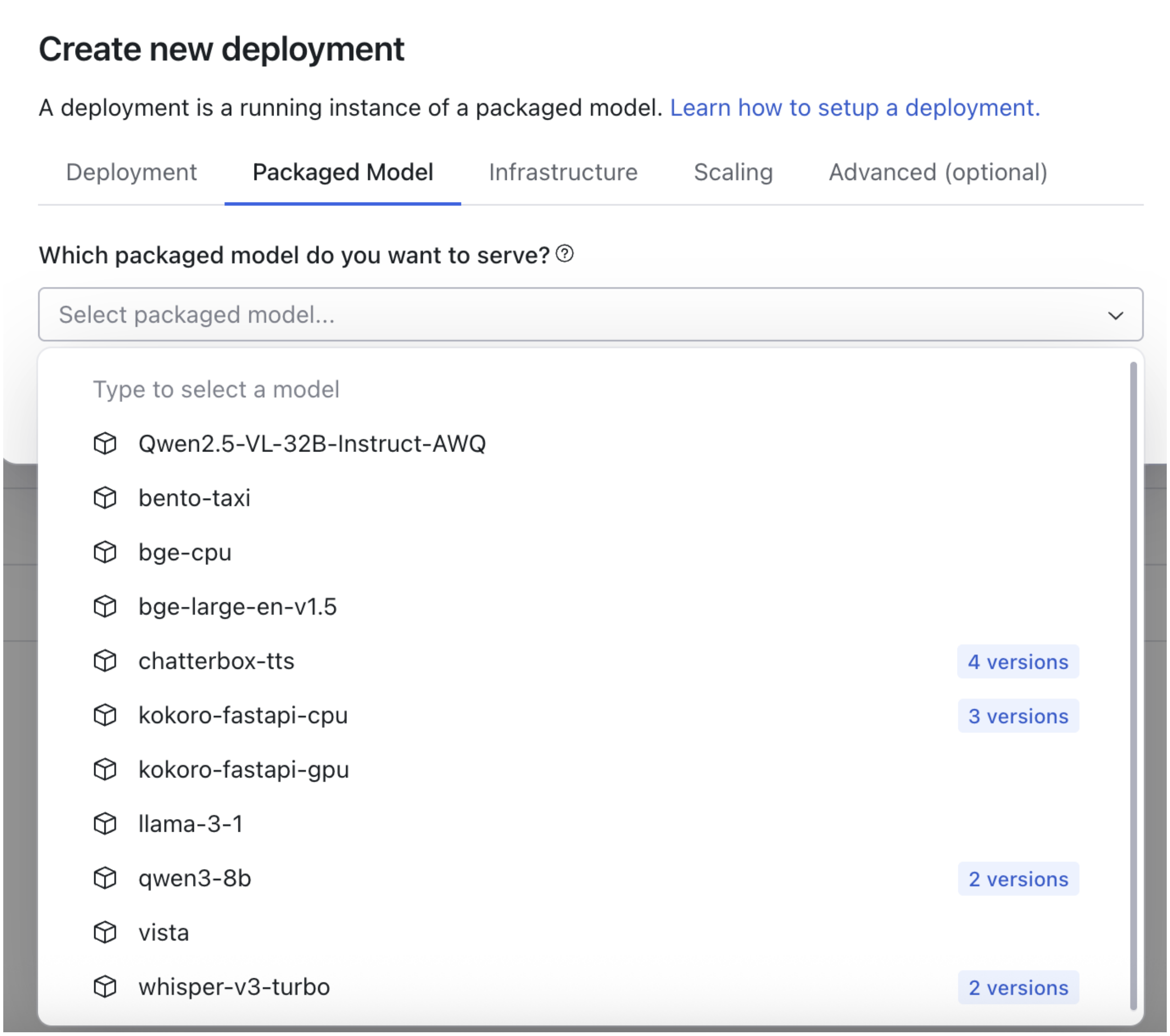
Create new deployment

A deployment is a running instance of a packaged model. Learn how to setup a deployment.

Deployment Name

vista-two

Select packaged model you just created.



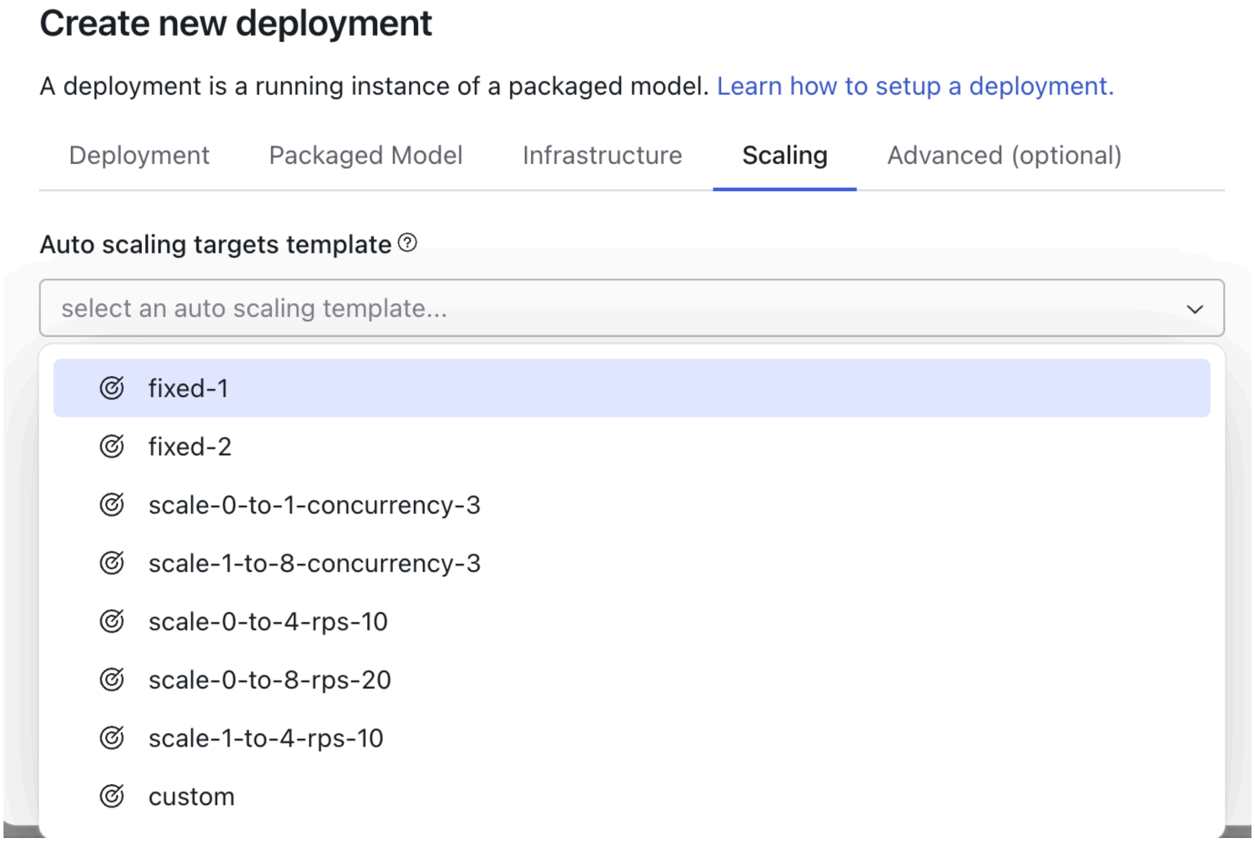
Create new deployment

A deployment is a running instance of a packaged model. Learn how to setup a deployment.

Which packaged model do you want to serve?

- Qwen2.5-VL-32B-Instruct-AWQ
- bento-taxi
- bge-cpu
- bge-large-en-v1.5
- chatterbox-tts 4 versions
- kokoro-fastapi-cpu 3 versions
- kokoro-fastapi-gpu
- llama-3-1
- qwen3-8b 2 versions
- vista
- whisper-v3-turbo 2 versions

next set auto scaling to fixed-1



Create new deployment

A deployment is a running instance of a packaged model. Learn how to setup a deployment.

Deployment | Packaged Model | Infrastructure | **Scaling** | Advanced (optional)

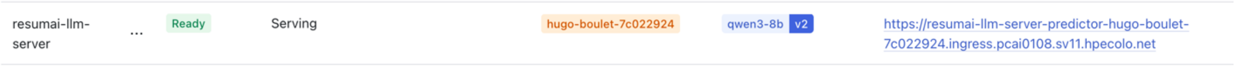
Auto scaling targets template

select an auto scaling template...

- fixed-1
- fixed-2
- scale-0-to-1-concurrency-3
- scale-1-to-8-concurrency-3
- scale-0-to-4-rps-10
- scale-0-to-8-rps-20
- scale-1-to-4-rps-10
- custom

Select Done, and wait for the model to deploy, this will take a few minutes.

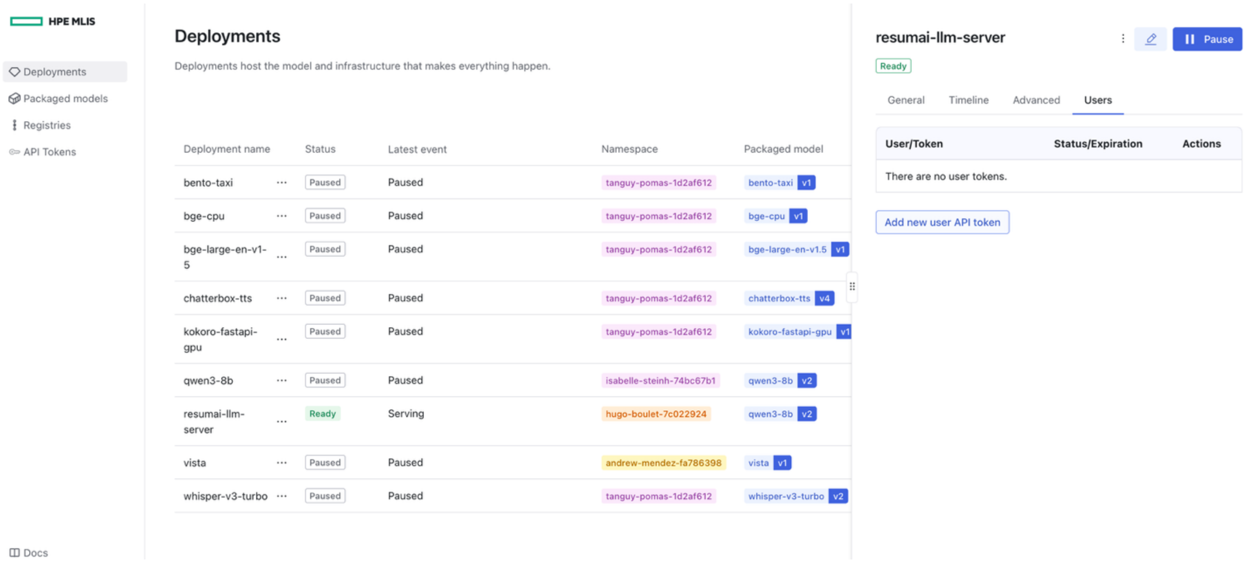
When its deployed, copy URL, example shown here



resumai-llm-server	...	Ready	Serving	hugo-boulet-7c022924	qwen3-8b v2	https://resumai-llm-server-predictor-hugo-boulet-7c022924.ingress.pcai0108.sv11.hpecolo.net

Next, lets make an API key

Select deployed deployment, go to users and click add new user api token



Deployments

Deployments host the model and infrastructure that makes everything happen.

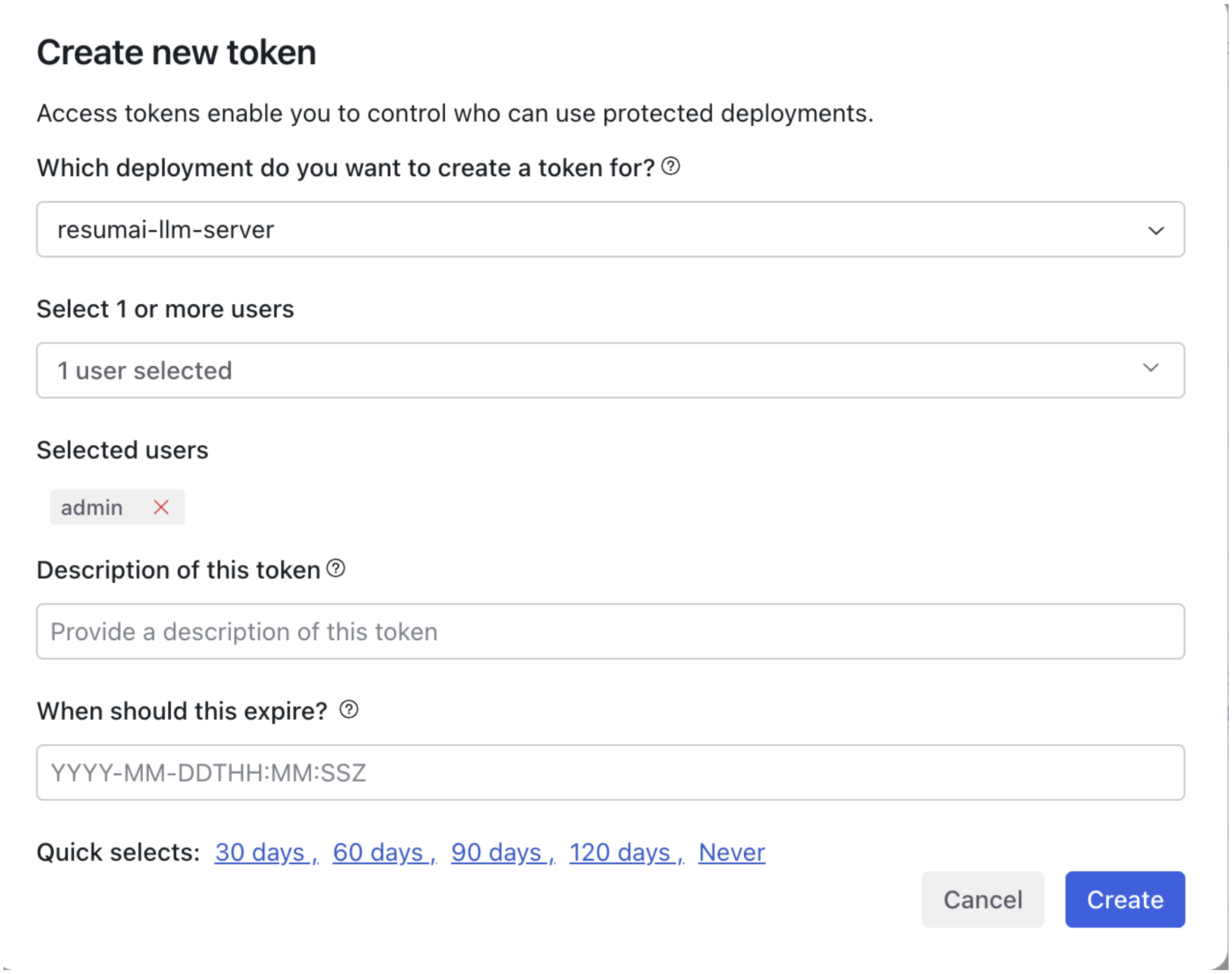
Deployment name	Status	Latest event	Namespace	Packaged model
bento-taxi	Paused	Paused	tanguy-pomas-1d2af612	bento-taxi v1
bge-cpu	Paused	Paused	tanguy-pomas-1d2af612	bge-cpu v1
bge-large-en-v1-5	Paused	Paused	tanguy-pomas-1d2af612	bge-large-en-v1.5 v1
chatterbox-tts	Paused	Paused	tanguy-pomas-1d2af612	chatterbox-tts v4
kokoro-fastapi-gpu	Paused	Paused	tanguy-pomas-1d2af612	kokoro-fastapi-gpu v1
qwen3-8b	Paused	Paused	isabelle-steinh-74bc67b1	qwen3-8b v2
resumai-llm-server	Ready	Serving	hugo-boulet-7c022924	qwen3-8b v2
vista	Paused	Paused	andrew-mendez-fa786398	vista v1
whisper-v3-turbo	Paused	Paused	tanguy-pomas-1d2af612	whisper-v3-turbo v2

resumai-llm-server

Ready

User/Token	Status/Expiration	Actions
There are no user tokens.		

Can set the role to whatever, I usually do admin



Create new token

Access tokens enable you to control who can use protected deployments.

Which deployment do you want to create a token for? ⓘ

resumai-llm-server

Select 1 or more users

1 user selected

Selected users

admin ✕

Description of this token ⓘ

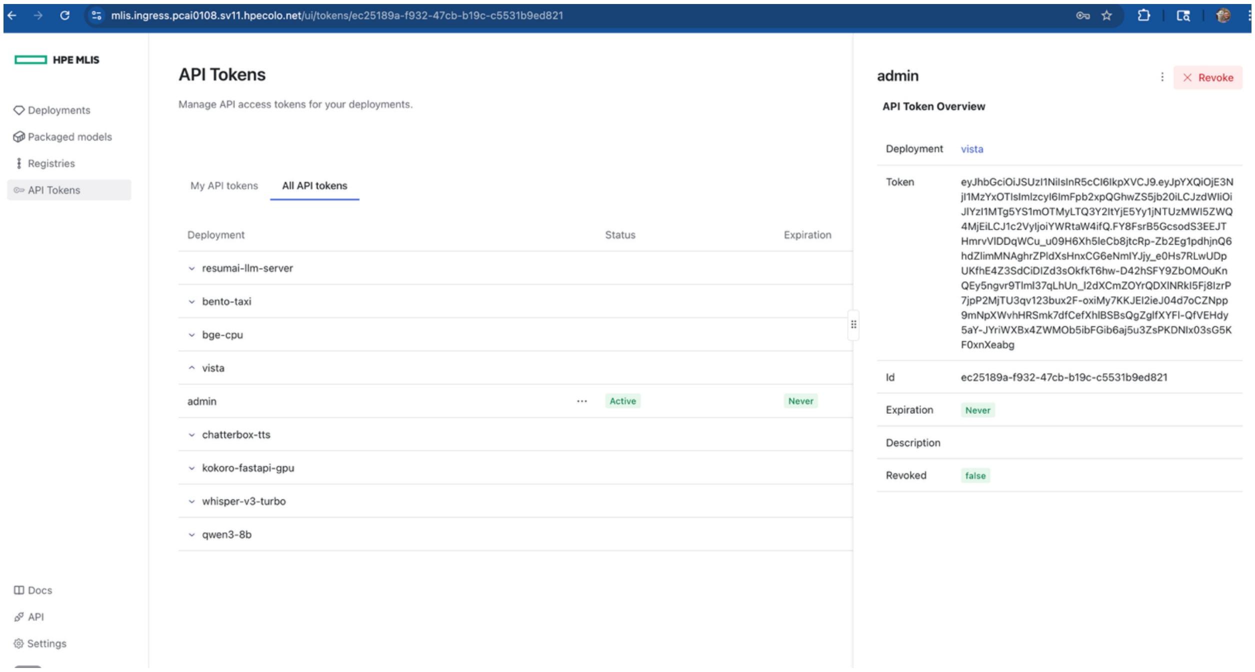
Provide a description of this token

When should this expire? ⓘ

YYYY-MM-DDTHH:MM:SSZ

Quick selects: 30 days , 60 days , 90 days , 120 days , Never

When the API is created, Select API Tokens, select all API tokens, and copy the API token shown on the right



mlis.ingress.pcai0108.sv11.hpecolo.net/ui/tokens/ec25189a-f932-47cb-b19c-c5531b9ed821

HPE MLIS

- Deployments
- Packaged models
- Registries
- API Tokens

API Tokens

Manage API access tokens for your deployments.

My API tokens | **All API tokens**

Deployment		Status	Expiration
⌄ resumai-llm-server			
⌄ bento-taxi			
⌄ bge-cpu			
⌃ vista			
admin	···	Active	Never
⌄ chatterbox-tts			
⌄ kokoro-fastapi-gpu			
⌄ whisper-v3-turbo			
⌄ qwen3-8b			

- Docs
- API
- Settings

admin ⋮ ✕ Revoke

API Token Overview

Deployment	vista
Token	eyJhbGciOiJSUzI1NiIsInR5cCI6IkpXVCJ9.eyJpYXQiOjE3NjI1MzYxOTIsImlzcyI6ImFpb2xpQGhwZS5jb20iLCJzdWIiOiJlYzI1MTg5YS1mOTMyLTQ3Y2ItYjE5Yy1jNTUzMWI5ZWQ4MjEiLCJ1c2VyIjoiYWRtaW4ifQ.FY8FsrB5GcsodS3EEJTHmrvVIDDqWCu_u09H6Xh5leCb8jtcRp-Zb2Eg1pdhjnQ6hdZlimMNAghrZPldXsHnxCG6eNmIYJjy_e0Hs7RLwUDpUKfhE4Z3SdCiDIZd3sOkfkT6hw-D42hSFY9ZbOMOuKnQEy5ngvr9Tlml37qLhUn_l2dXCmZOYrQDXlNRkI5Fj8IzrP7jpP2MjTU3qv123bux2F-oxiMy7KKJEl2ieJ04d7oCZNpp9mNpXWvhHRSmk7dfCefXhlBSBsQgZglfXYFl-QfVEHdy5aY-JYriWXBx4ZWMOb5ibFGib6aj5u3ZsPKDNlx03sG5KF0xnXeabg
Id	ec25189a-f932-47cb-b19c-c5531b9ed821
Expiration	Never
Description	
Revoked	false

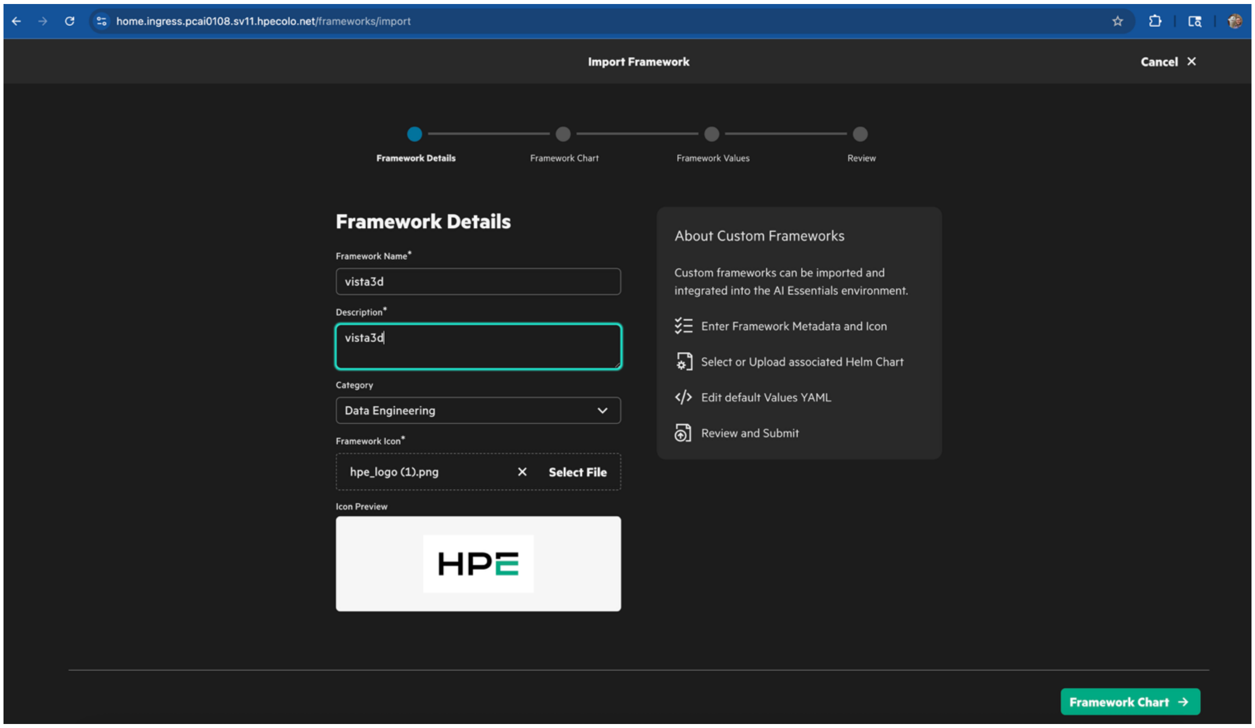
You are ready to deploy the helm chart

Pre-req; you will need the tgz file of the helm chart ready to upload.

Part 2: Install Helm chart

GO to Tools & Frameworks > Import Framework

Add name, description, and logo



Framework Details

Framework Name*
vista3d

Description*
vista3d

Category
Data Engineering

Framework Icon*
hpe_logo (1).png ✕

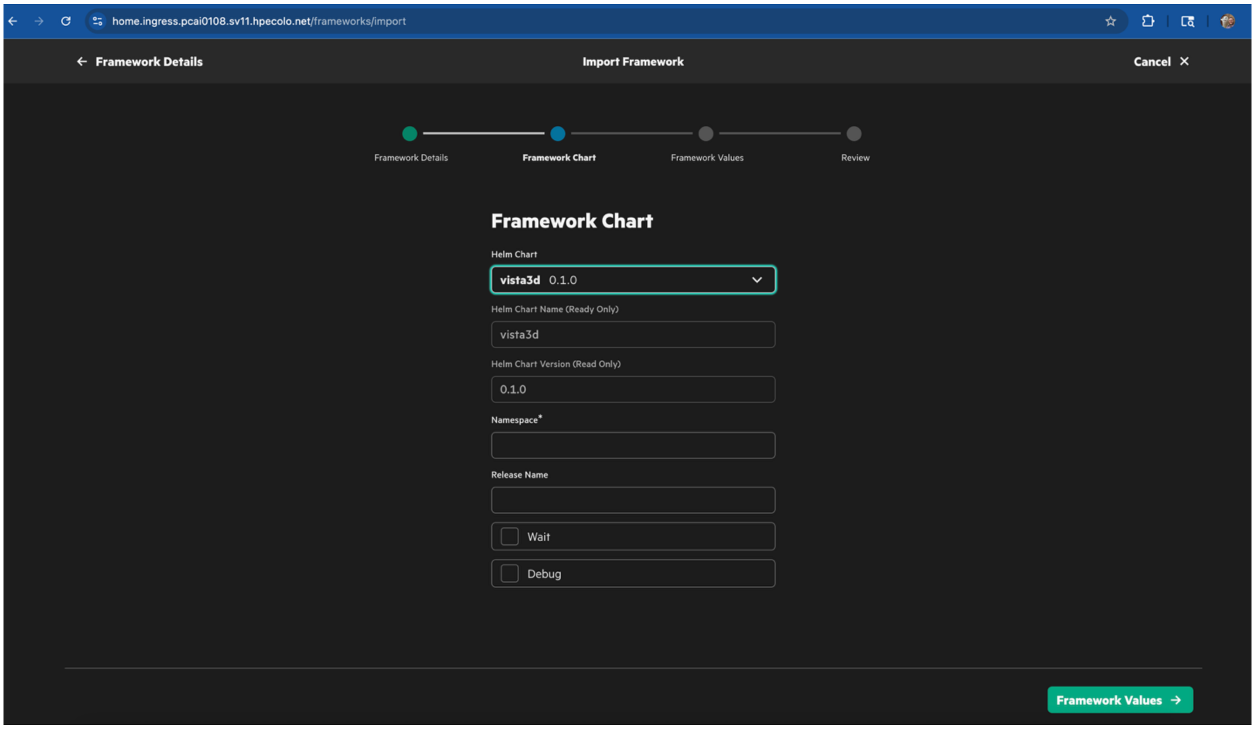
Icon Preview

About Custom Frameworks

Custom frameworks can be imported and integrated into the AI Essentials environment.

- Enter Framework Metadata and Icon
- Select or Upload associated Helm Chart
- Edit default Values YAML
- Review and Submit

Drag .tgz of helm chart in UI, you should see this next



Framework Chart

Helm Chart
vista3d 0.1.0

Helm Chart Name (Ready Only)
vista3d

Helm Chart Version (Read Only)
0.1.0

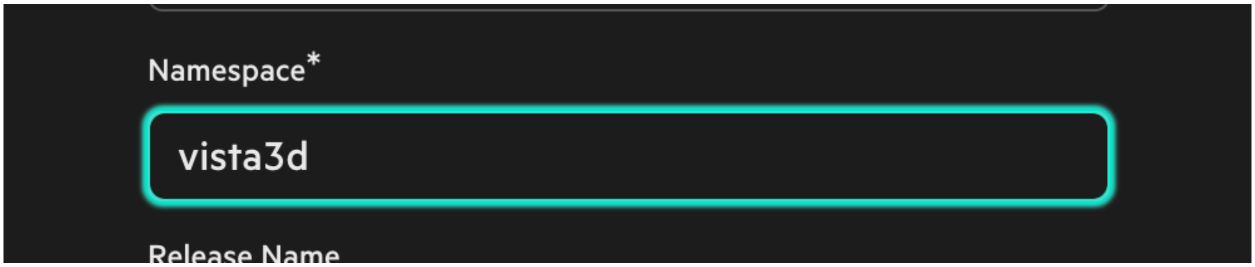
Namespace*

Release Name

□ Wait

□ Debug

Add namespace

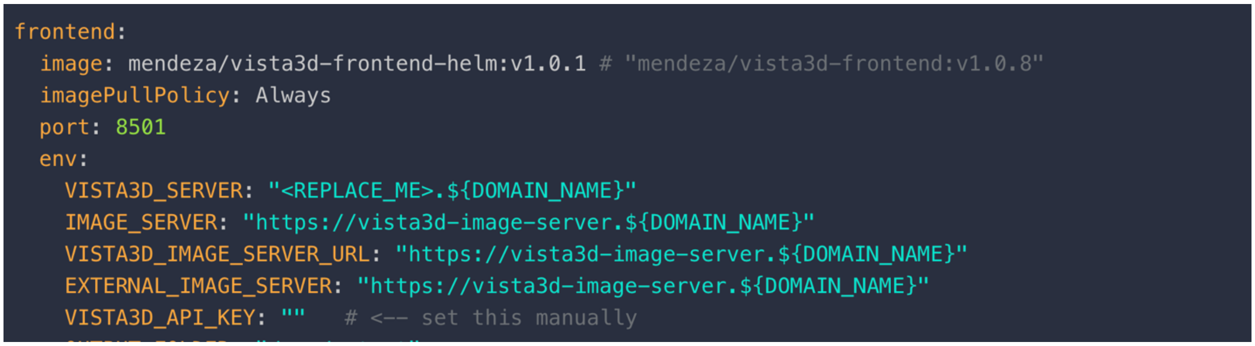


Namespace*

vista3d

Release Name

Now in the helm chart UI, only change the VISTA3D_SERVER and the VISTA3D_API_KEY



```
frontend:
  image: mendeza/vista3d-frontend-helm:v1.0.1 # "mendeza/vista3d-frontend:v1.0.8"
  imagePullPolicy: Always
  port: 8501
  env:
    VISTA3D_SERVER: "<REPLACE_ME>.${DOMAIN_NAME}"
    IMAGE_SERVER: "https://vista3d-image-server.${DOMAIN_NAME}"
    VISTA3D_IMAGE_SERVER_URL: "https://vista3d-image-server.${DOMAIN_NAME}"
    EXTERNAL_IMAGE_SERVER: "https://vista3d-image-server.${DOMAIN_NAME}"
    VISTA3D_API_KEY: ""   # <-- set this manually
```

Wait until its deployed, if its deployed successfully, you should see Open Button.