How to add models the pre-deployed NIM model to MLIS

Create NGC Registry

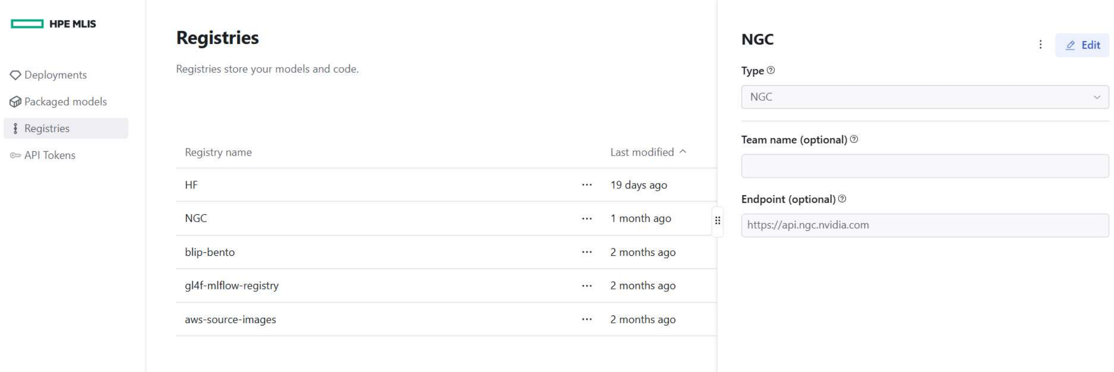
Ideally to access NIM models you will need to have a NGC key.

But these NIM models under Tools & Frameworks >> NVIDIA AI Enterprise, are already available on the system for use.

There is an in-built registry which is valid only for these models which come along with the system.

Thus, you do not need to register for NGC key on MLIS for these models.

Do verify that there is an entry under Registries as show below.



Registries

Registries store your models and code.

Registry name	Last modified
HF	19 days ago
NGC	1 month ago
blip-bento	2 months ago
gl4f-mlflow-registry	2 months ago
aws-source-images	2 months ago

NGC

Type: NGC

Team name (optional):

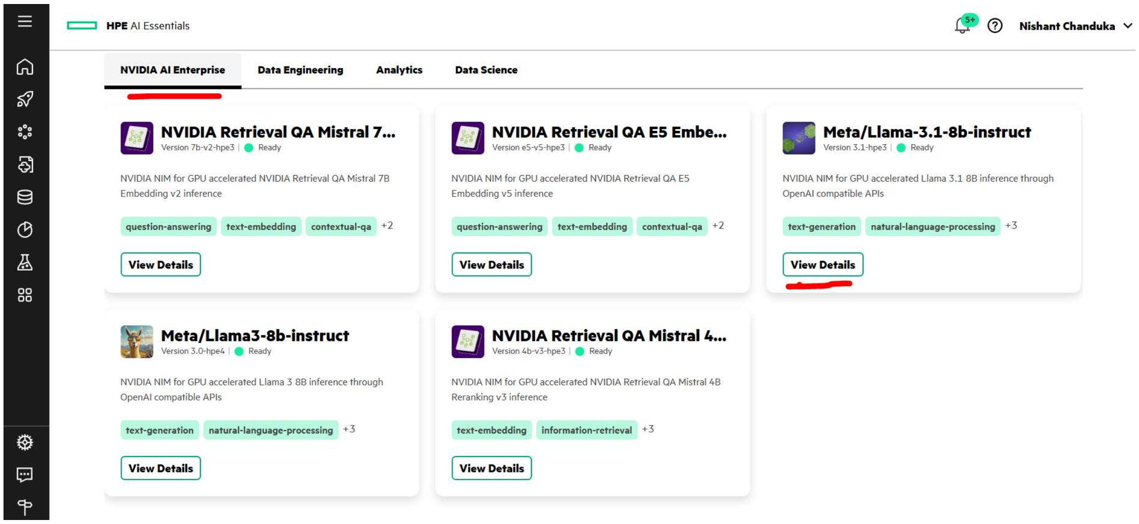
Endpoint (optional): https://api.ngc.nvidia.com

Packaged Model

Go to the NVIDIA AI Enterprise Tab.

Select the model you want to deploy.

Click on **View Details** for the model.



NVIDIA Retrieval QA Mistral 7...
Version 7b-v2-hpe3 | Ready
NVIDIA NIM for GPU accelerated NVIDIA Retrieval QA Mistral 7B Embedding v2 inference
question-answering text-embedding contextual-qa +2
View Details

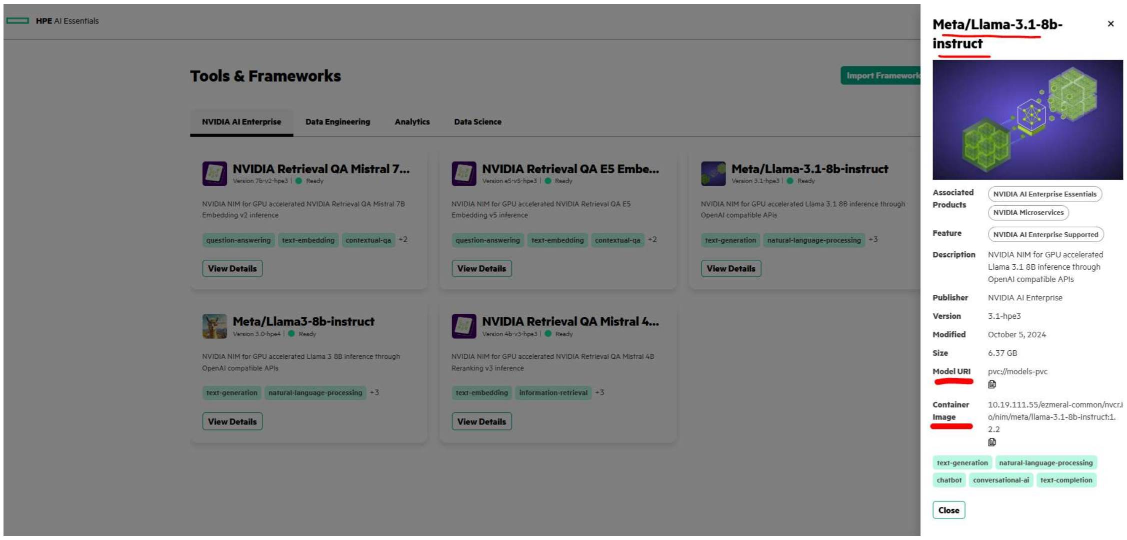
NVIDIA Retrieval QA E5 Embe...
Version e5-v5-hpe3 | Ready
NVIDIA NIM for GPU accelerated NVIDIA Retrieval QA E5 Embedding v5 inference
question-answering text-embedding contextual-qa +2
View Details

Meta/Llama-3.1-8b-instruct
Version 3.1-hpe3 | Ready
NVIDIA NIM for GPU accelerated Llama 3.1 8B inference through OpenAI compatible APIs
text-generation natural-language-processing +3
View Details

Meta/Llama3-8b-instruct
Version 3.0-hpe4 | Ready
NVIDIA NIM for GPU accelerated Llama 3 8B inference through OpenAI compatible APIs
text-generation natural-language-processing +3
View Details

NVIDIA Retrieval QA Mistral 4...
Version 4b-v3-hpe3 | Ready
NVIDIA NIM for GPU accelerated NVIDIA Retrieval QA Mistral 4B Reranking v3 inference
text-embedding information-retrieval +3
View Details

Make note of the parameters marked in red as they will be required while packaging the model for being hosted on MLIS.



Tools & Frameworks

NVIDIA Retrieval QA Mistral 7...
Version 7b-v2-hpe3 | Ready
NVIDIA NIM for GPU accelerated NVIDIA Retrieval QA Mistral 7B Embedding v2 inference
question-answering text-embedding contextual-qa +2

NVIDIA Retrieval QA E5 Embe...
Version e5-v5-hpe3 | Ready
NVIDIA NIM for GPU accelerated NVIDIA Retrieval QA E5 Embedding v5 inference
question-answering text-embedding contextual-qa +2

Meta/Llama-3.1-8b-instruct
Version 3.1-hpe3 | Ready
NVIDIA NIM for GPU accelerated Llama 3.1 8B inference through OpenAI compatible APIs
text-generation natural-language-processing +3

Meta/Llama3-8b-instruct
Version 3.0-hpe4 | Ready
NVIDIA NIM for GPU accelerated Llama 3 8B inference through OpenAI compatible APIs
text-generation natural-language-processing +3

NVIDIA Retrieval QA Mistral 4...
Version 4b-v3-hpe3 | Ready
NVIDIA NIM for GPU accelerated NVIDIA Retrieval QA Mistral 4B Reranking v3 inference
text-embedding information-retrieval +3

Meta/Llama-3.1-8b-instruct

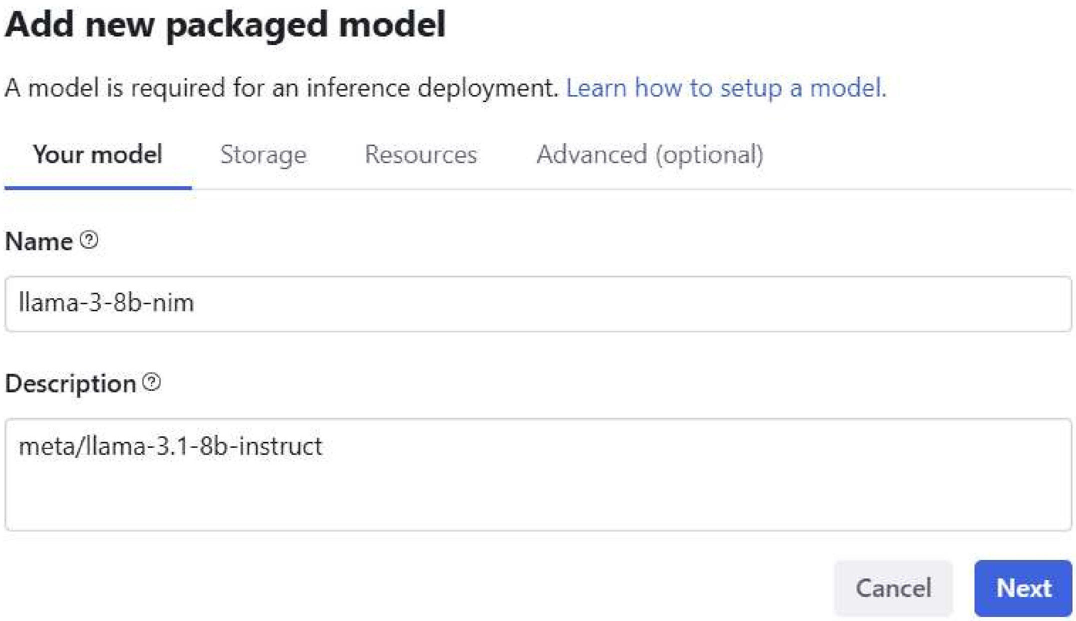
Associated Products	NVIDIA AI Enterprise Essentials NVIDIA Microservices
Feature	NVIDIA AI Enterprise Supported
Description	NVIDIA NIM for GPU accelerated Llama 3.1 8B inference through OpenAI compatible APIs
Publisher	NVIDIA AI Enterprise
Version	3.1-hpe3
Modified	October 5, 2024
Size	6.37 GB
Model URI	pvc://models-pvc
Container Image	10.19.111.55/ezmeral-common/nvcr.io/nim/meta/llama-3.1-8b-instruct:1.2.2

text-generation natural-language-processing chatbot conversational-ai text-completion

Now back to the MLIS page.

Name: Enter a name which helps you to identify the model easily, preferably model name.

Description: Favourably add the actual model's name. So, its easy for reference later.



Add new packaged model

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

Name ⓘ

llama-3-8b-nim

Description ⓘ

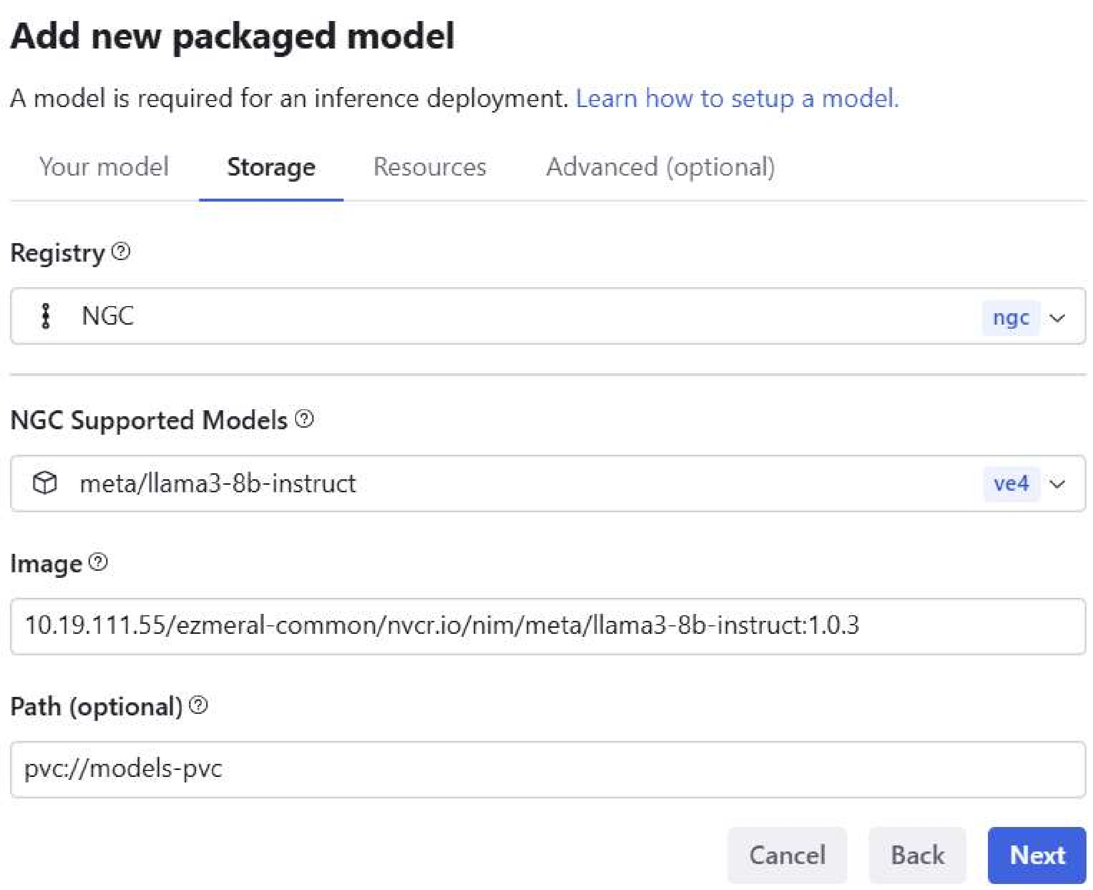
meta/llama-3.1-8b-instruct

In the Storage tab,

Registry: Select the **NGC** registry.

As soon as you select that the **NGC Supported Models will pe pre-populated with values:** Select the model you want to deploy from the drop down.

As soon as you select the model from the drop down, the fields **Image** and **Path** will be auto filled. The values should match the values from the model's **View Details** page captured earlier.



Add new packaged model

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

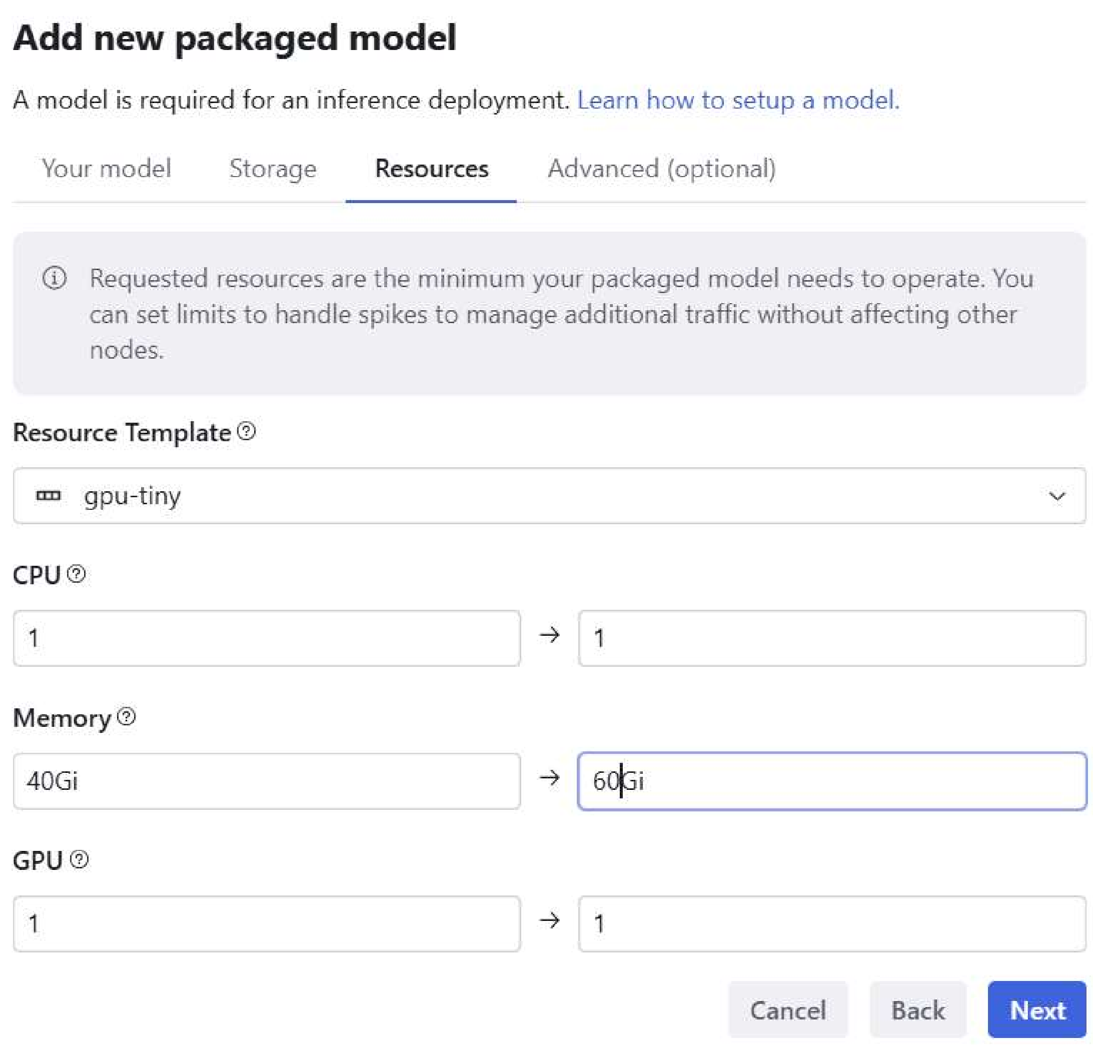
Registry: NGC (ngc)

NGC Supported Models: meta/llama3-8b-instruct (ve4)

Image: 10.19.111.55/ezmeral-common/nvcr.io/nim/meta/llama3-8b-instruct:1.0.3

Path (optional): pvc://models-pvc

Resource Template: Choose appropriate template or define a custom one based on the model requirements. Refer the documentation for help.



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-tiny

CPU: 1 → 1

Memory: 40Gi → 60Gi

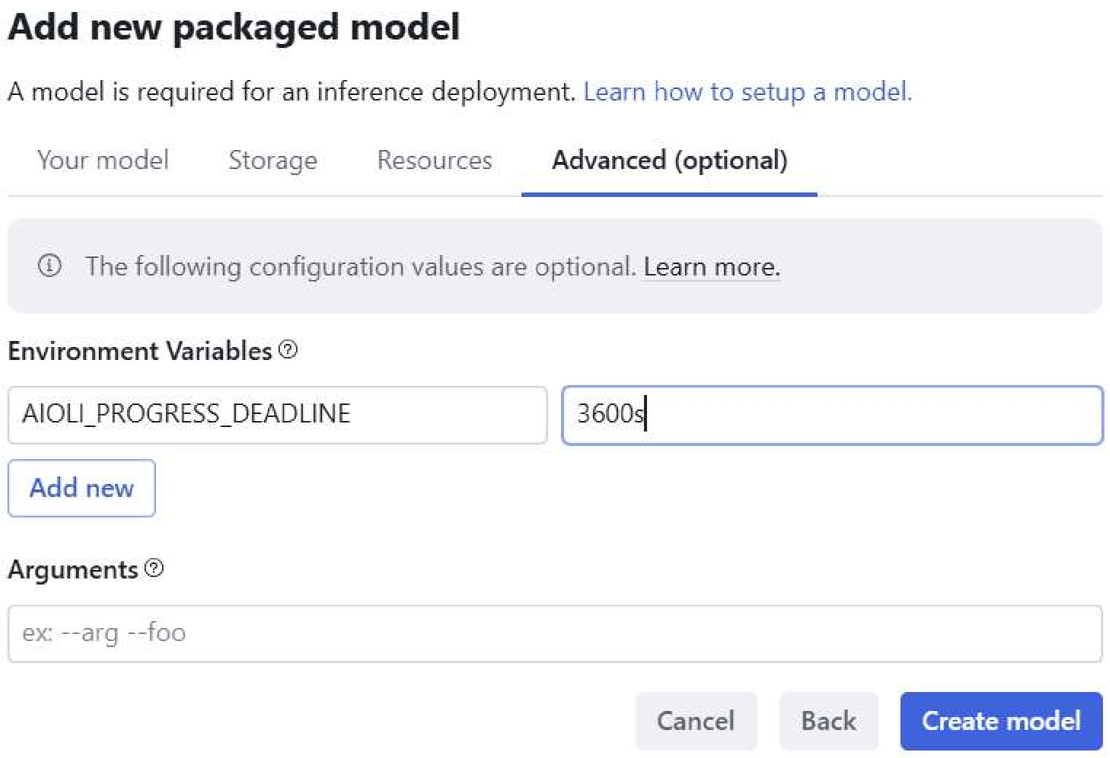
GPU: 1 → 1

Advanced (optional)

Environment Variables: Add the required ones. Refer documentation

AIOLI_PROGRESS_DEADLINE: 3600s

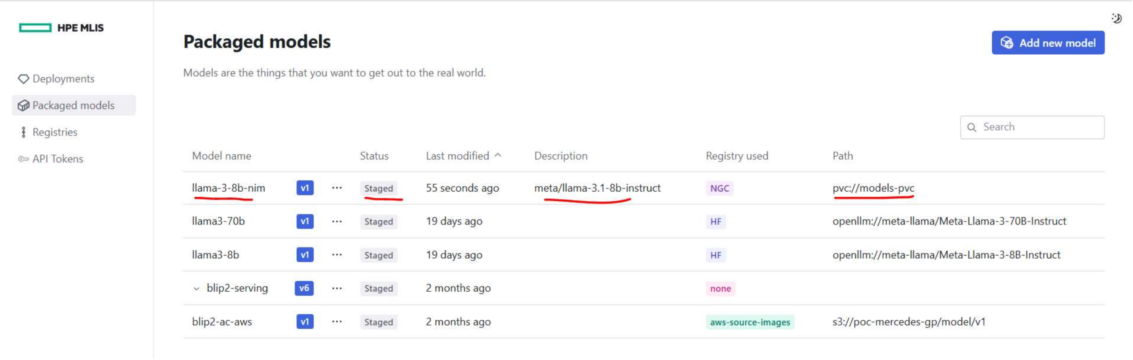
The above indicates deadline for downloading the model: default value is 30 mins. Thus, it is most commonly set since for some models this can go beyond that time frame.


Add new packaged model
A model is required for an inference deployment. Learn how to setup a model.
Your model
Storage
Resources
Advanced (optional)
The following configuration values are optional. Learn more.
Add new
Cancel
Back
Create model

Environment Variables

AIOLI_PROGRESS_DEADLINE: 3600s

Arguments: ex: --arg --foo



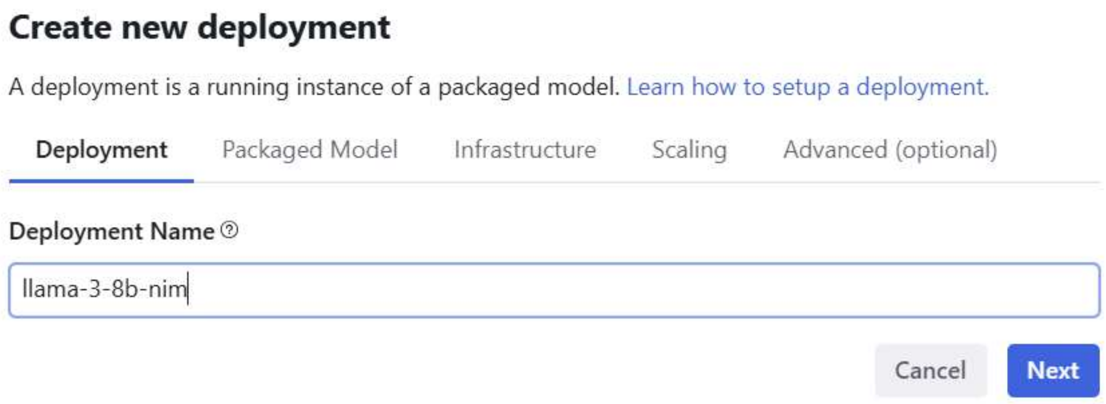
Packaged models

Models are the things that you want to get out to the real world.

Model name			Status	Last modified ^	Description	Registry used	Path
llama-3-8b-nim	v1	···	Staged	55 seconds ago	meta/llama-3.1-8b-instruct	NGC	pvc://models-pvc
llama3-70b	v1	···	Staged	19 days ago		HF	openllm://meta-llama/Meta-Llama-3-70B-Instruct
llama3-8b	v1	···	Staged	19 days ago		HF	openllm://meta-llama/Meta-Llama-3-8B-Instruct
⌄ blip2-serving	v6	···	Staged	2 months ago		none	
blip2-ac-aws	v1	···	Staged	2 months ago		aws-source-images	s3://poc-mercedes-gp/model/v1

Deployments

Deployment Name: Provide a name of your choice. just ensure it needs to be all in lowercase and you can only use special characters "." and "-".

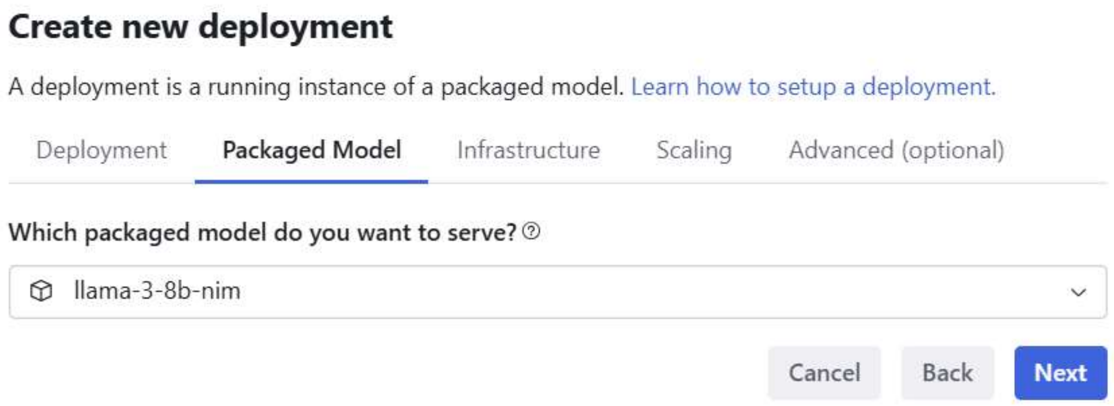


Create new deployment

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

Deployment Name: llama-3-8b-nim

Packaged Model: Select the packaged model created in previous step.

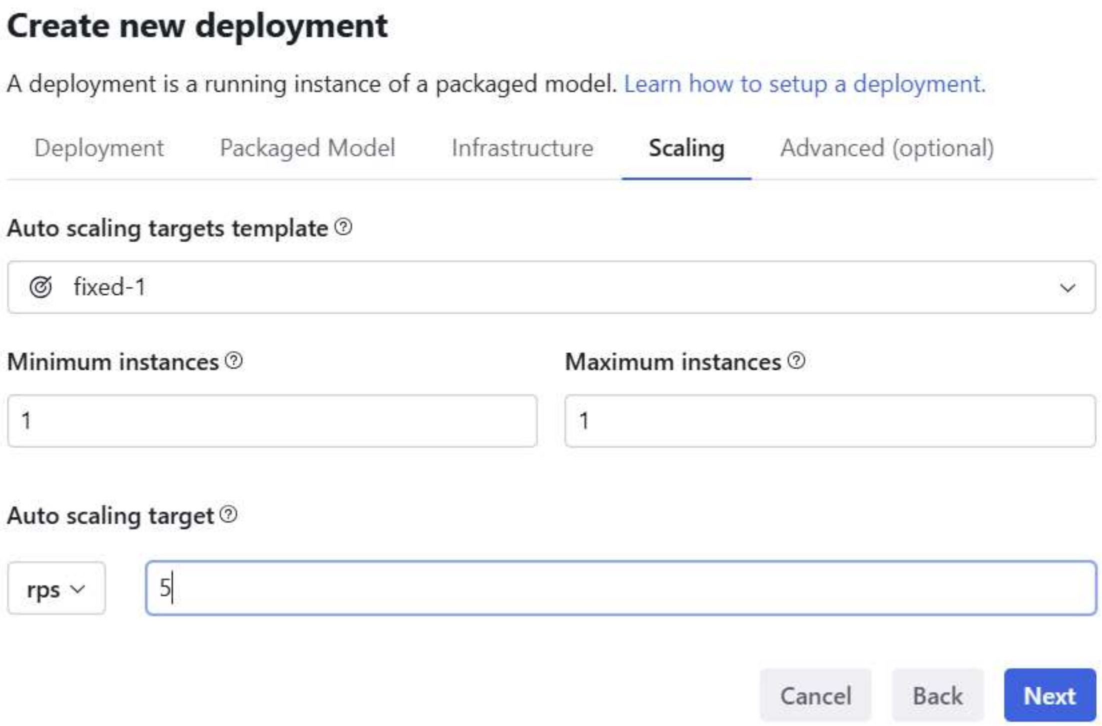


Create new deployment

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

Which packaged model do you want to serve?: llama-3-8b-nim

Scaling: Choose option as applicable, for testing purpose we go with fixed-1.



Create new deployment

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

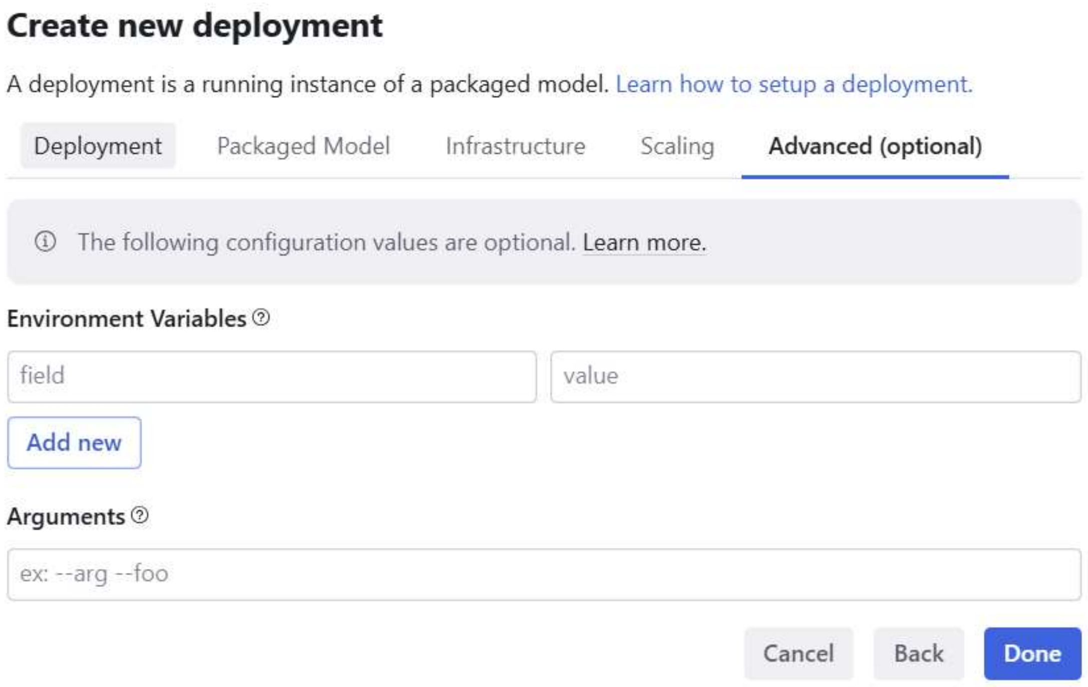
Auto scaling targets template: fixed-1

Minimum instances: 1

Maximum instances: 1

Auto scaling target: rps 5

Advanced: Fill in as applicable (normally left blank)



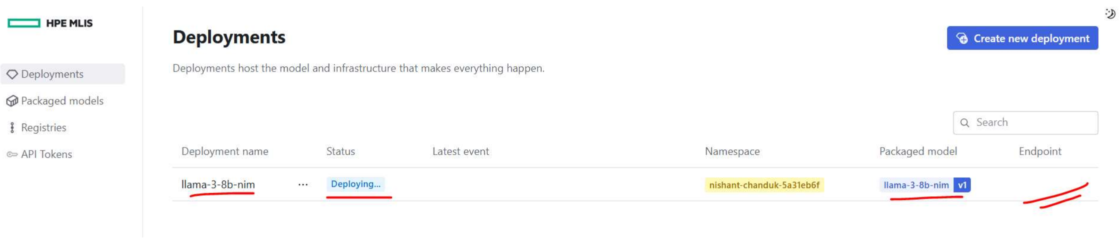
Create new deployment

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

The following configuration values are optional. Learn more.

Environment Variables

Arguments



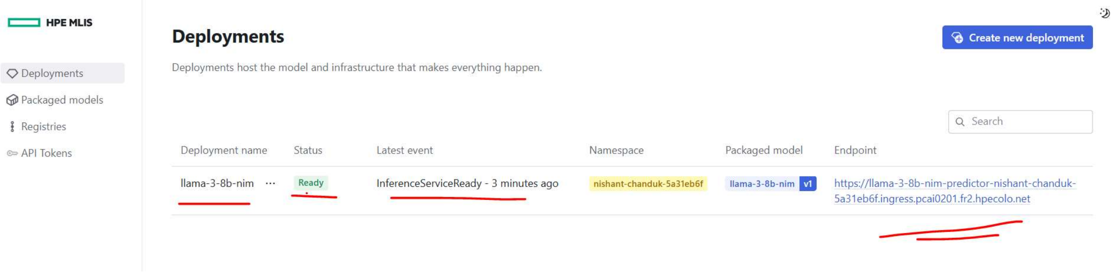
Deployments

Deployments host the model and infrastructure that makes everything happen.

Deployment name	Status	Latest event	Namespace	Packaged model	Endpoint
llama-3-8b-nim ···	Deploying...		nishant-chanduk-5a31eb6f	llama-3-8b-nim v1	

The above snapshot shows the model is getting deployed which will take some time.

Once the model is in **Ready** state you would see a valid endpoint.



Deployments

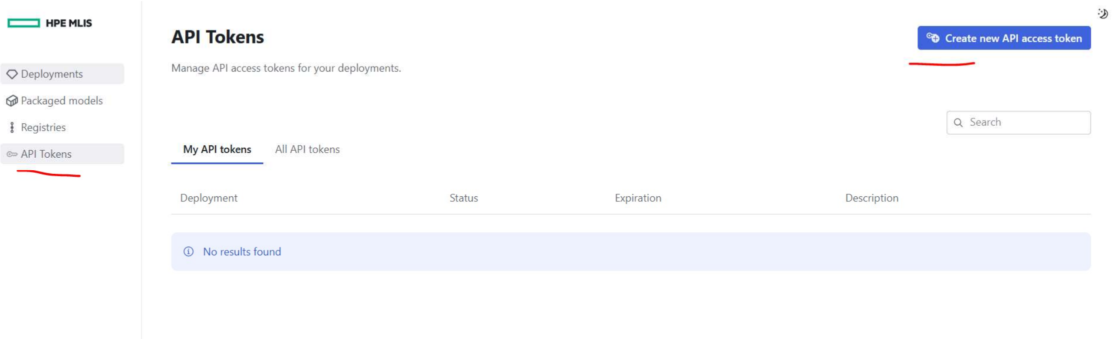
Deployments host the model and infrastructure that makes everything happen.

Deployment name	Status	Latest event	Namespace	Packaged model	Endpoint
llama-3-8b-nim ···	Ready	InferenceServiceReady - 3 minutes ago	nishant-chanduk-5a31eb6f	llama-3-8b-nim v1	https://llama-3-8b-nim-predictor-nishant-chanduk-5a31eb6f.ingress.pcai0201.fr2.hpecolo.net

Now how to we access the model anywhere within or outside the system.

Go to **API Tokens**

Click on **Create new API access Token**



API Tokens

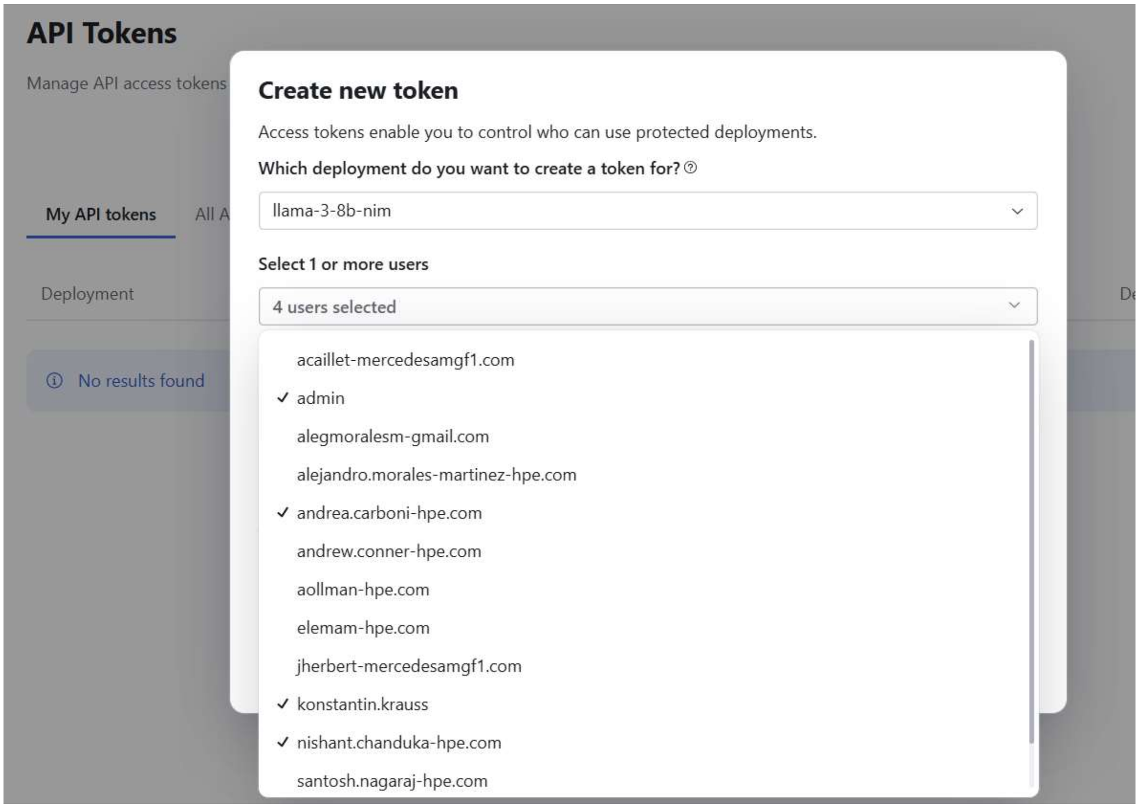
Manage API access tokens for your deployments.

Deployment	Status	Expiration	Description
No results found			

The dropdown menu **Which deployment do you want to create a token for?** Will list the models you see in the **Deployments** tab only.

Select your model.

Select 1 or more users: Select the users you want to be able to use the token.



Create new token

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

Which deployment do you want to create a token for?

llama-3-8b-nim

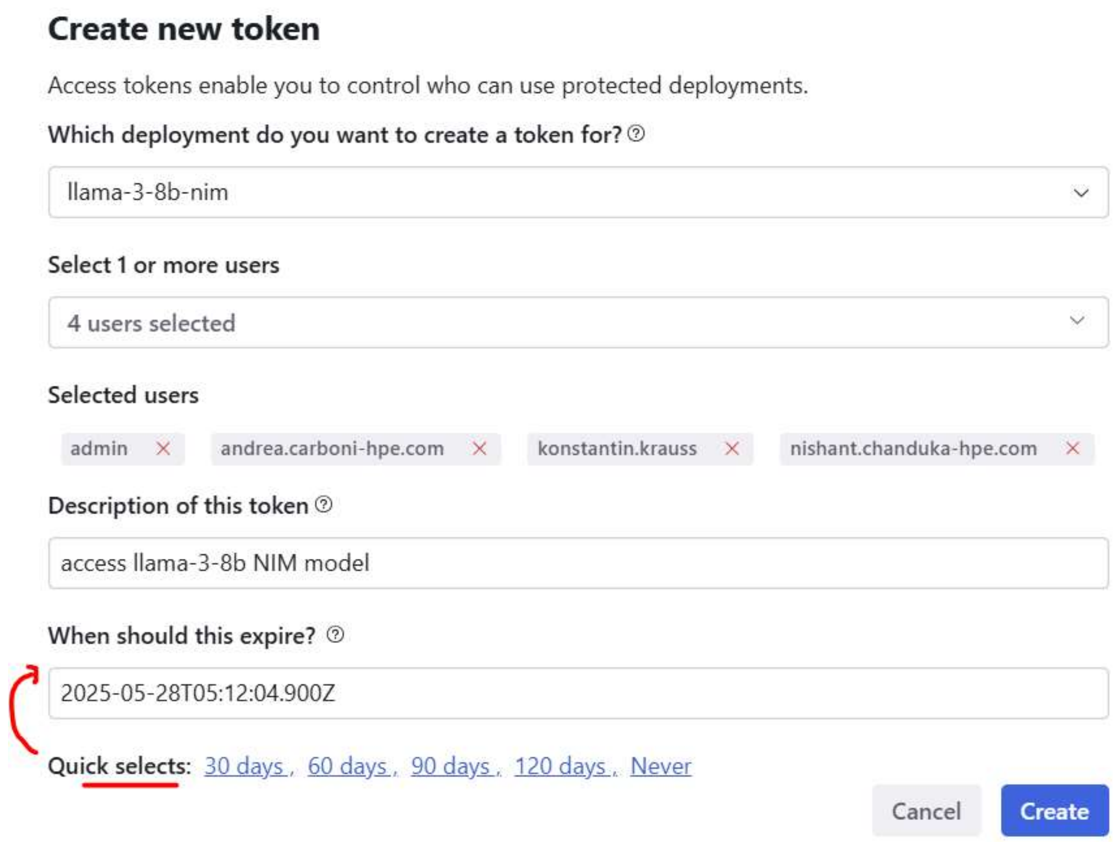
Select 1 or more users

4 users selected

acaillet-mercedesamgf1.com
✓ admin
alegmoralesm-gmail.com
alejandro.morales-martinez-hpe.com
✓ andrea.carboni-hpe.com
andrew.conner-hpe.com
aollman-hpe.com
elemam-hpe.com
jherbert-mercedesamgf1.com
✓ konstantin.krauss
✓ nishant.chanduka-hpe.com
santosh.nagaraj-hpe.com

Add **Description of this token**

Select **When should this expire?**



Create new token

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

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

llama-3-8b-nim

Select 1 or more users

4 users selected

Selected users

admin ✕ andrea.carboni-hpe.com ✕ konstantin.krauss ✕ nishant.chanduka-hpe.com ✕

Description of this token ⓘ

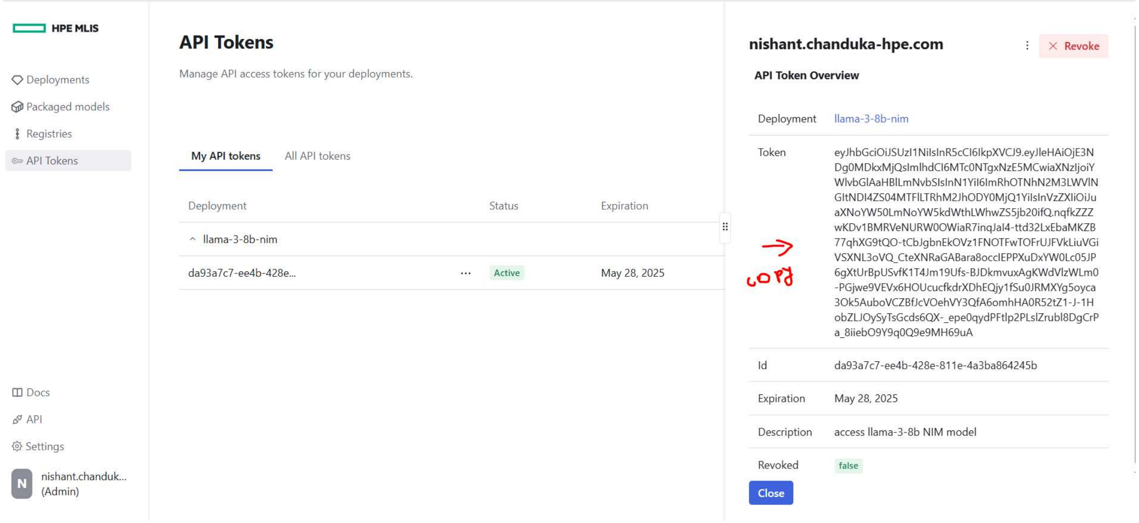
access llama-3-8b NIM model

When should this expire? ⓘ

2025-05-28T05:12:04.900Z

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

Copy the Created token



API Tokens

Manage API access tokens for your deployments.

My API tokens | All API tokens

Deployment		Status	Expiration
^ llama-3-8b-nim			
da93a7c7-ee4b-428e...	···	Active	May 28, 2025

nishant.chanduka-hpe.com

API Token Overview

Deployment	llama-3-8b-nim
Token	eyJhbGciOiJSUzI1NiIsInR5cCI6IkpXVCJ9.eyJleHAiOjE3NDg0MDkxMjQsImlhdCI6MTc0NTgxNzE5MCwiaXNzIjoiYWlvbGlAaHBlLmNvbSIsInN1YiI6ImRhOTNhN2M3LWVlNGItNDI4ZS04MTFlLTRhM2JhODY0MjQ1YiIsInVzZXIiOiJuaXNoYW50LmNoYW5kdWthLWhwZS5jb20ifQ.nqfkZZZwKDv1BMRVeNURW0OWiaR7inqJaI4-ttd32LxEbaMKZB77qhXG9tQO-tCbJgbnEkOVz1FNOTFwTOFrUJFVkLiuVGiVSXNL3oVQ_CteXNRaGABara8occIEPPXuDxYW0Lc05JP6gXtUrBpUSvfK1T4Jm19Ufs-BJDkmvuxAgKWdVlzWLm0-PGjwe9VEVx6HOUcucfkdrXDhEQjy1fSu0JRMXYg5oyca3Ok5AuboVCZBfJcVOehVY3QfA6omhHA0R52tZ1-J-1HobZLJOySyTsGcds6QX-_epe0qydPFtlp2PLslZrubl8DgCrPa_8iiebO9Y9q0Q9e9MH69uA
Id	da93a7c7-ee4b-428e-811e-4a3ba864245b
Expiration	May 28, 2025
Description	access llama-3-8b NIM model
Revoked	false

copy