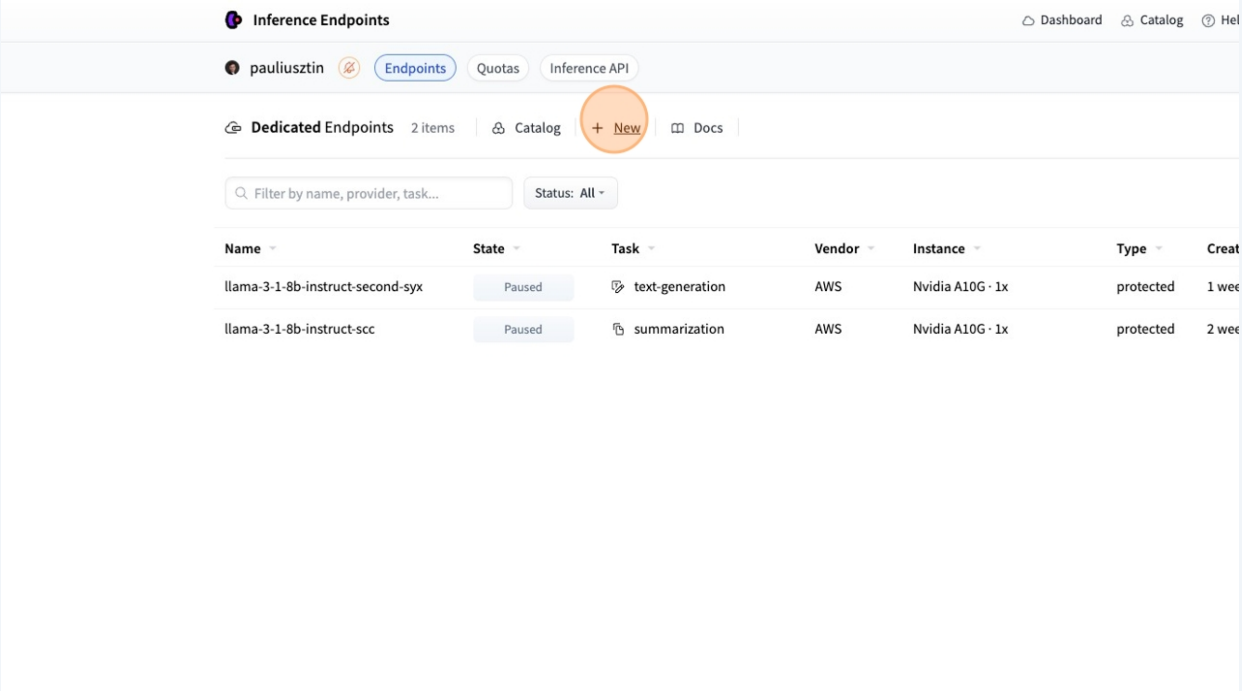
Creating an Inference Endpoint on Hugging Face

1 Click "New"



Inference Endpoints

Dashboard · Catalog · Hel

pauliusztin · Endpoints · Quotas · Inference API

Dedicated Endpoints 2 items · Catalog · + New · Docs

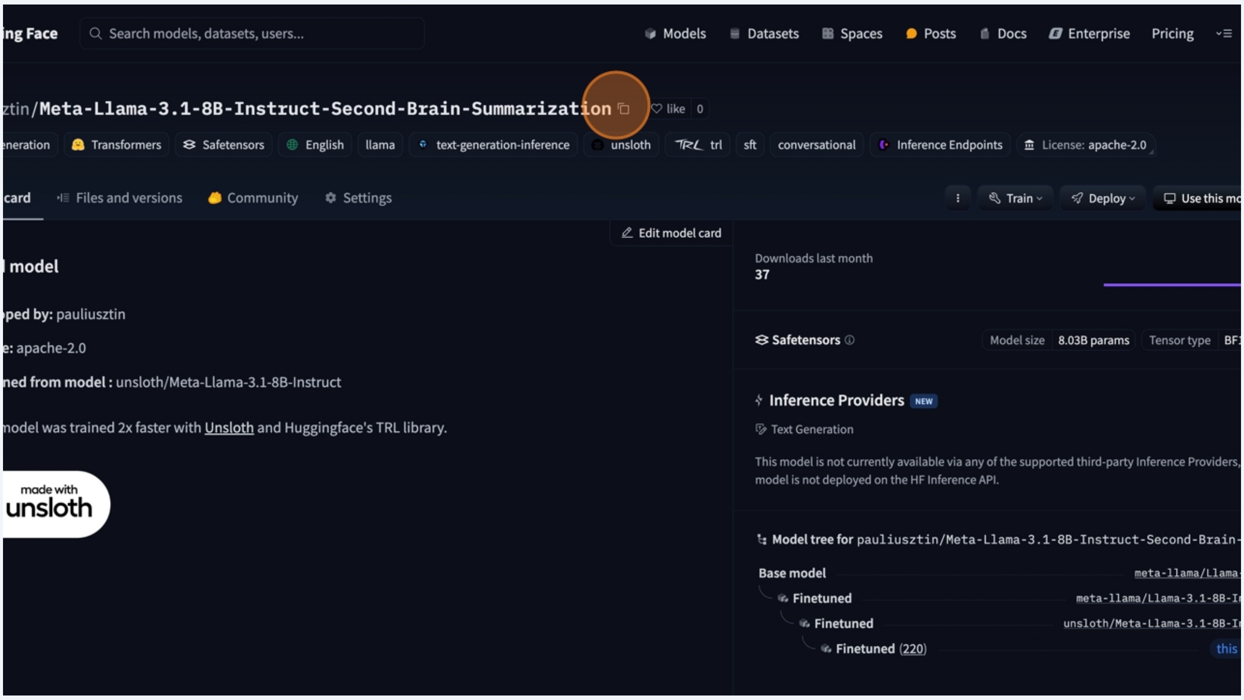
Filter by name, provider, task...

Status: All

Name	State	Task	Vendor	Instance	Type	Creat
llama-3-1-8b-instruct-second-syx	Paused	text-generation	AWS	Nvidia A10G · 1x	protected	1 wee
llama-3-1-8b-instruct-scc	Paused	summarization	AWS	Nvidia A10G · 1x	protected	2 wee

2 Switch to tab pauliusztin/Meta-Llama-3.1-8B-Instruct-Second-Brain-Summarization · Hugging Face"

3 Click this icon.



ztin/Meta-Llama-3.1-8B-Instruct-Second-Brain-Summarization

l model

ped by: pauliusztin

e: apache-2.0

ned from model : unsloth/Meta-Llama-3.1-8B-Instruct

model was trained 2x faster with Unsloth and Huggingface's TRL library.

This model is not currently available via any of the supported third-party Inference Providers, model is not deployed on the HF Inference API.

Model tree for pauliusztin/Meta-Llama-3.1-8B-Instruct-Second-Brain-

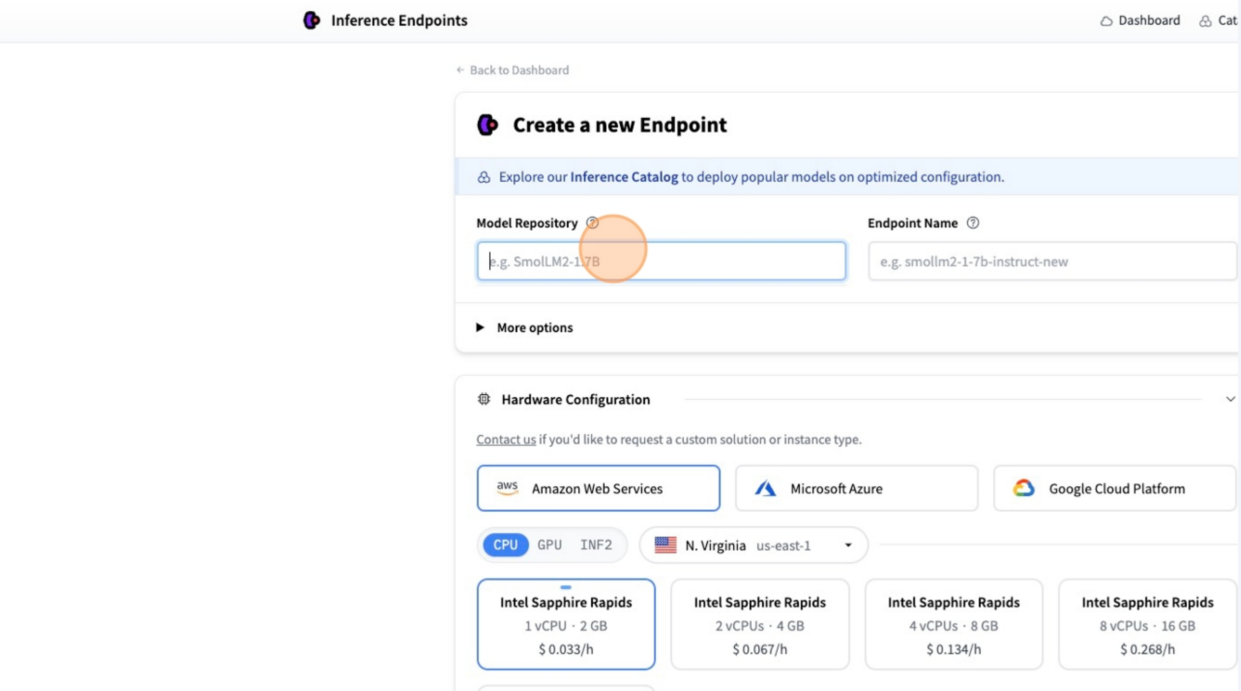
meta-llama/Llama-

meta-llama/Llama-3.1-8B-I

unsloth/Meta-Llama-3.1-8B-I

4 Switch to tab Create a new Endpoint | Inference Endpoints by Hugging Face"

5 Enter the Hugging Face model ID you copied in the "Model Repository" field.



Create a new Endpoint

Explore our **Inference Catalog** to deploy popular models on optimized configuration.

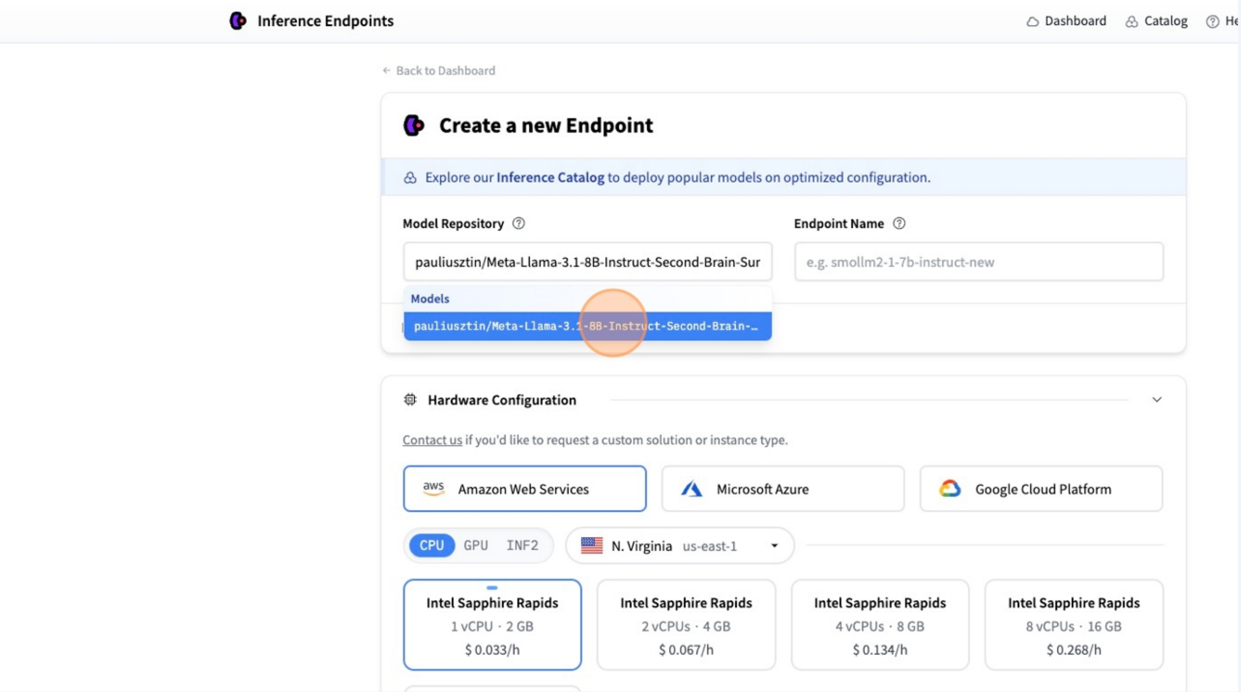
Model Repository: e.g. SmolLM2-1.7B

Endpoint Name: e.g. smollm2-1-7b-instruct-new

Contact us if you'd like to request a custom solution or instance type.

Intel Sapphire Rapids 1 vCPU · 2 GB $ 0.033/h	Intel Sapphire Rapids 2 vCPUs · 4 GB $ 0.067/h	Intel Sapphire Rapids 4 vCPUs · 8 GB $ 0.134/h	Intel Sapphire Rapids 8 vCPUs · 16 GB $ 0.268/h

6 We used "pauliusztin/Meta-Llama-3.1-8B-Instruct-Second-Brain-Summarization":



Create a new Endpoint

Explore our **Inference Catalog** to deploy popular models on optimized configuration.

Model Repository: pauliusztin/Meta-Llama-3.1-8B-Instruct-Second-Brain-Sur

Models

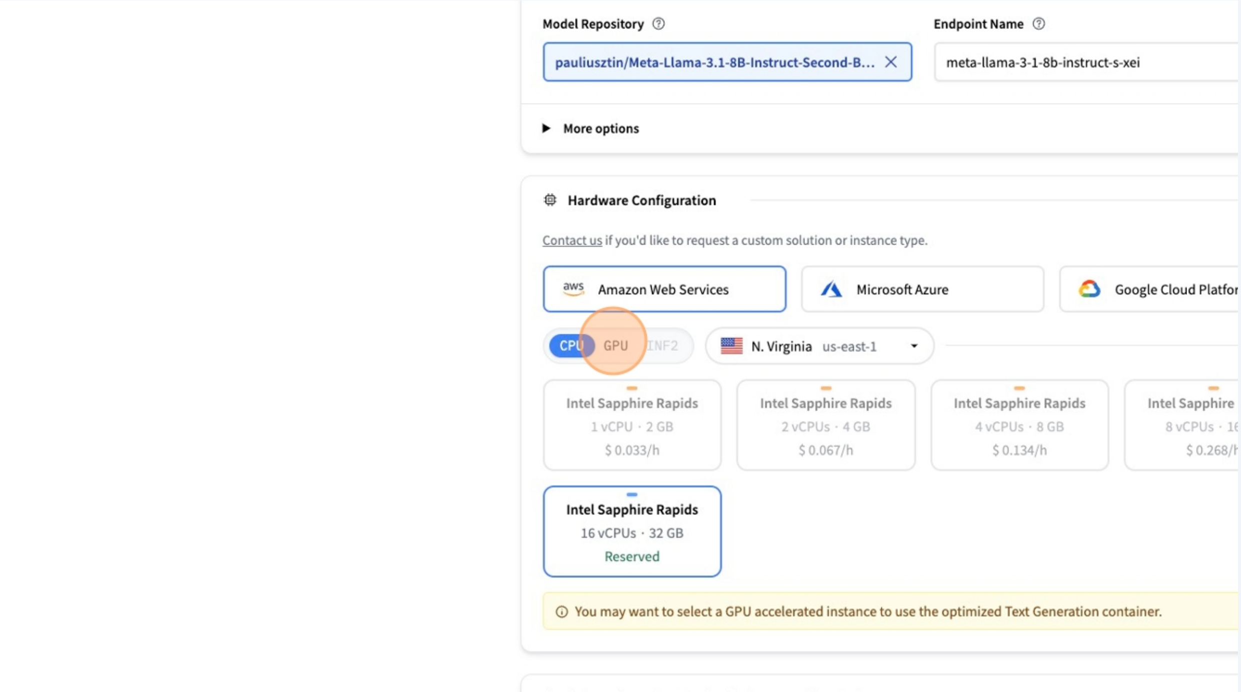
pauliusztin/Meta-Llama-3.1-8B-Instruct-Second-Brain-…

Endpoint Name: e.g. smollm2-1-7b-instruct-new

Contact us if you'd like to request a custom solution or instance type.

Intel Sapphire Rapids 1 vCPU · 2 GB $ 0.033/h	Intel Sapphire Rapids 2 vCPUs · 4 GB $ 0.067/h	Intel Sapphire Rapids 4 vCPUs · 8 GB $ 0.134/h	Intel Sapphire Rapids 8 vCPUs · 16 GB $ 0.268/h

7 Click "GPU"



Model Repository: pauliusztin/Meta-Llama-3.1-8B-Instruct-Second-B...

Endpoint Name: meta-llama-3-1-8b-instruct-s-xei

▶ More options

Hardware Configuration

Contact us if you'd like to request a custom solution or instance type.

Amazon Web Services | Microsoft Azure | Google Cloud Platfor

CPU | GPU | INF2 — N. Virginia us-east-1

Intel Sapphire Rapids — 1 vCPU · 2 GB — $ 0.033/h

Intel Sapphire Rapids — 2 vCPUs · 4 GB — $ 0.067/h

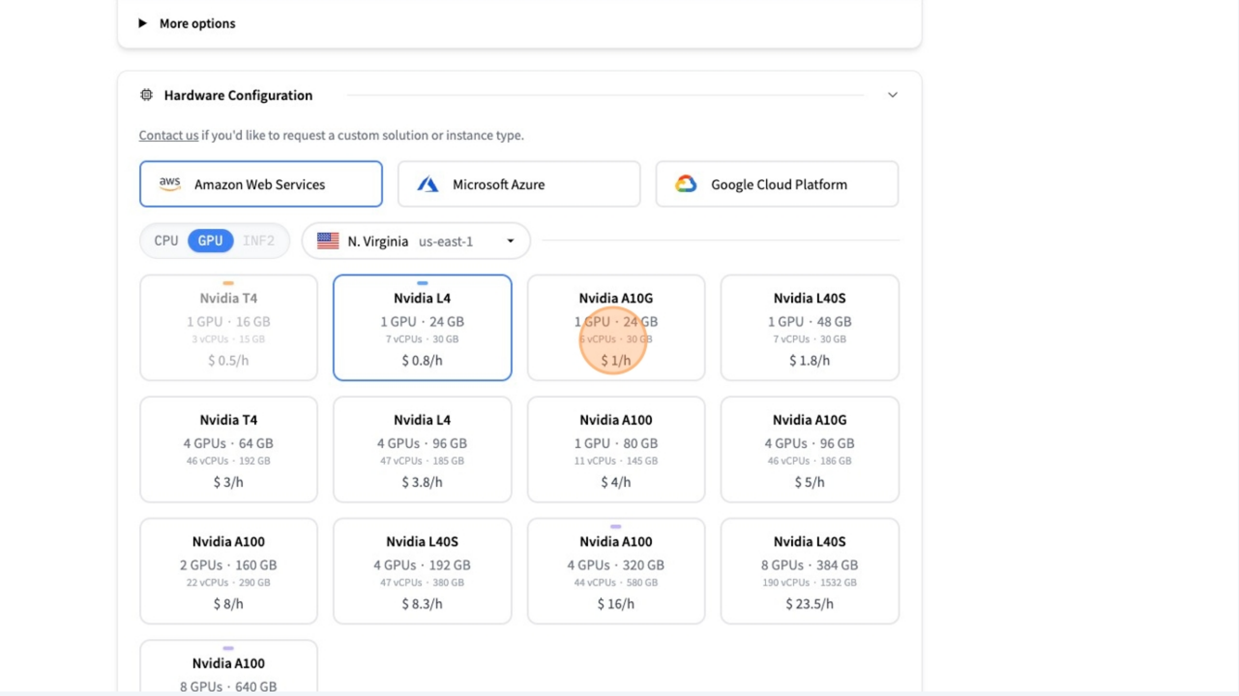
Intel Sapphire Rapids — 4 vCPUs · 8 GB — $ 0.134/h

Intel Sapphire — 8 vCPUs · 16 — $ 0.268/h

Intel Sapphire Rapids — 16 vCPUs · 32 GB — Reserved

ⓘ You may want to select a GPU accelerated instance to use the optimized Text Generation container.

8 Choose an Nvidia A10G GPU.



▶ More options

Hardware Configuration

Contact us if you'd like to request a custom solution or instance type.

Amazon Web Services | Microsoft Azure | Google Cloud Platform

CPU | GPU | INF2 — N. Virginia us-east-1

Nvidia T4 — 1 GPU · 16 GB — 3 vCPUs · 15 GB — $ 0.5/h

Nvidia L4 — 1 GPU · 24 GB — 7 vCPUs · 30 GB — $ 0.8/h

Nvidia A10G — 1 GPU · 24 GB — 6 vCPUs · 30 GB — $ 1/h

Nvidia L40S — 1 GPU · 48 GB — 7 vCPUs · 30 GB — $ 1.8/h

Nvidia T4 — 4 GPUs · 64 GB — 46 vCPUs · 192 GB — $ 3/h

Nvidia L4 — 4 GPUs · 96 GB — 47 vCPUs · 185 GB — $ 3.8/h

Nvidia A100 — 1 GPU · 80 GB — 11 vCPUs · 145 GB — $ 4/h

Nvidia A10G — 4 GPUs · 96 GB — 46 vCPUs · 186 GB — $ 5/h

Nvidia A100 — 2 GPUs · 160 GB — 22 vCPUs · 290 GB — $ 8/h

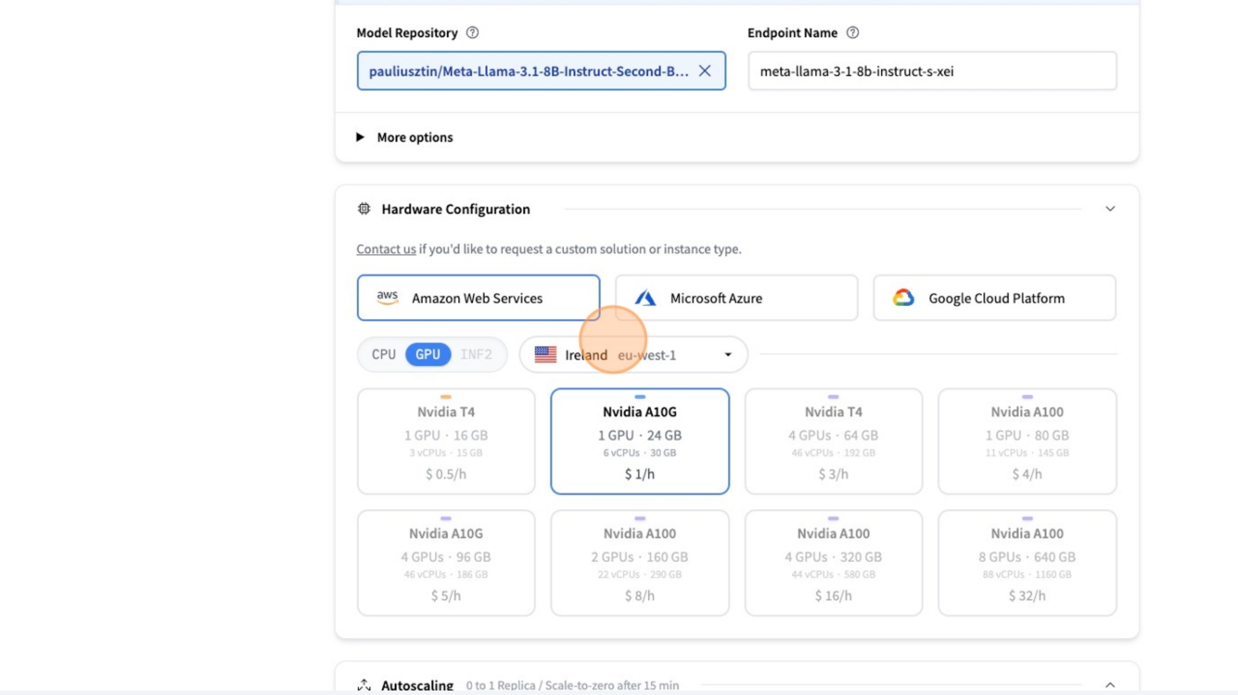
Nvidia L40S — 4 GPUs · 192 GB — 47 vCPUs · 380 GB — $ 8.3/h

Nvidia A100 — 4 GPUs · 320 GB — 44 vCPUs · 580 GB — $ 16/h

Nvidia L40S — 8 GPUs · 384 GB — 190 vCPUs · 1532 GB — $ 23.5/h

Nvidia A100 — 8 GPUs · 640 GB

9 Select the closest region to you, such as "Ireland [eu-west-1]"



Model Repository: pauliusztin/Meta-Llama-3.1-8B-Instruct-Second-B...

Endpoint Name: meta-llama-3-1-8b-instruct-s-xei

Hardware Configuration

Contact us if you'd like to request a custom solution or instance type.

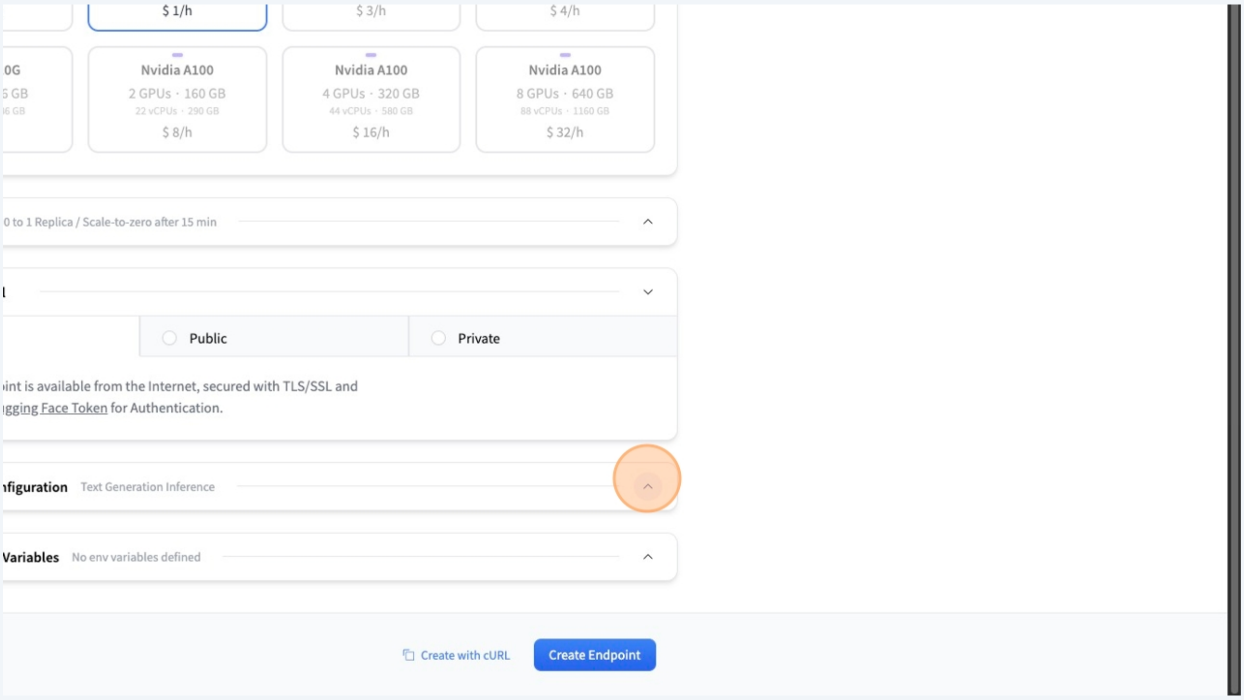
Amazon Web Services | Microsoft Azure | Google Cloud Platform

CPU | GPU | INF2 — Ireland eu-west-1

Nvidia T4 1 GPU · 16 GB 3 vCPUs · 15 GB $ 0.5/h	Nvidia A10G 1 GPU · 24 GB 6 vCPUs · 30 GB $ 1/h	Nvidia T4 4 GPUs · 64 GB 46 vCPUs · 192 GB $ 3/h	Nvidia A100 1 GPU · 80 GB 11 vCPUs · 145 GB $ 4/h
Nvidia A10G 4 GPUs · 96 GB 46 vCPUs · 186 GB $ 5/h	Nvidia A100 2 GPUs · 160 GB 22 vCPUs · 290 GB $ 8/h	Nvidia A100 4 GPUs · 320 GB 44 vCPUs · 580 GB $ 16/h	Nvidia A100 8 GPUs · 640 GB 88 vCPUs · 1160 GB $ 32/h

Autoscaling 0 to 1 Replica / Scale-to-zero after 15 min

10 Go to the "Configuration" section



0 to 1 Replica / Scale-to-zero after 15 min

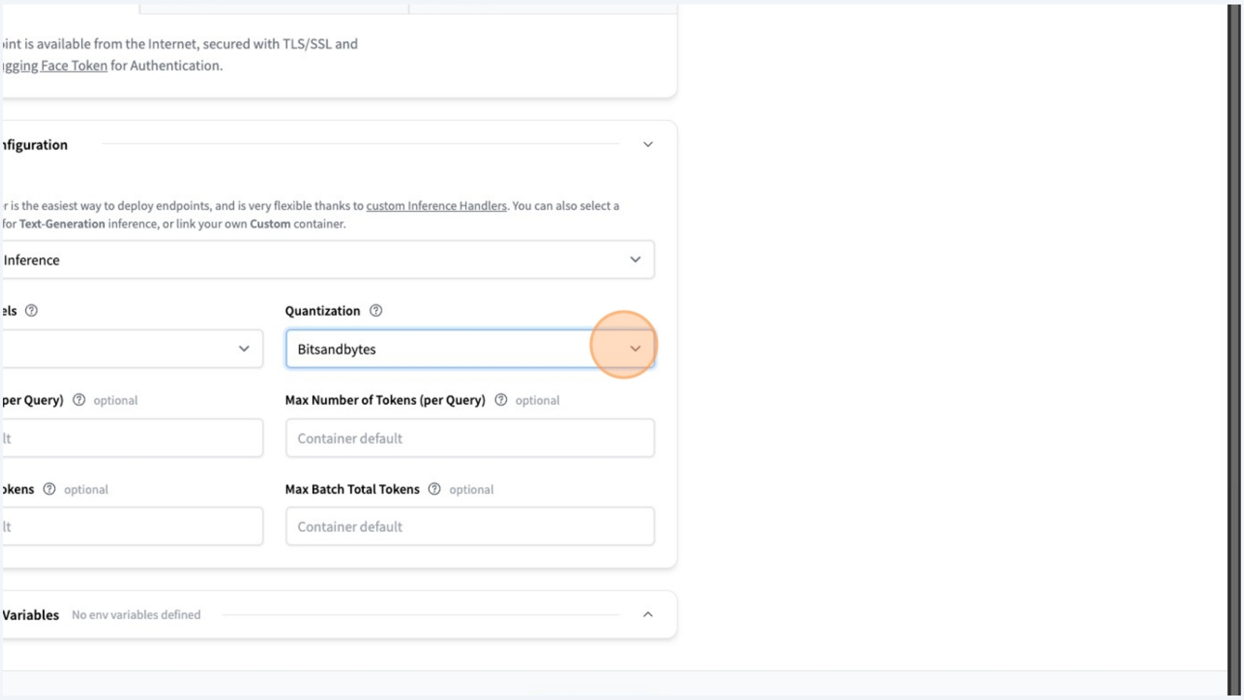
○ Public ○ Private

int is available from the Internet, secured with TLS/SSL and
gging Face Token for Authentication.

nfiguration Text Generation Inference

Variables No env variables defined

11 Select the "Bitsandbytes" quantization option



ɿint is available from the Internet, secured with TLS/SSL and
ɿgging Face Token for Authentication.

ɿfiguration

r is the easiest way to deploy endpoints, and is very flexible thanks to custom Inference Handlers. You can also select a for Text-Generation inference, or link your own Custom container.

Inference

ɿels

Quantization: Bitsandbytes

per Query) optional

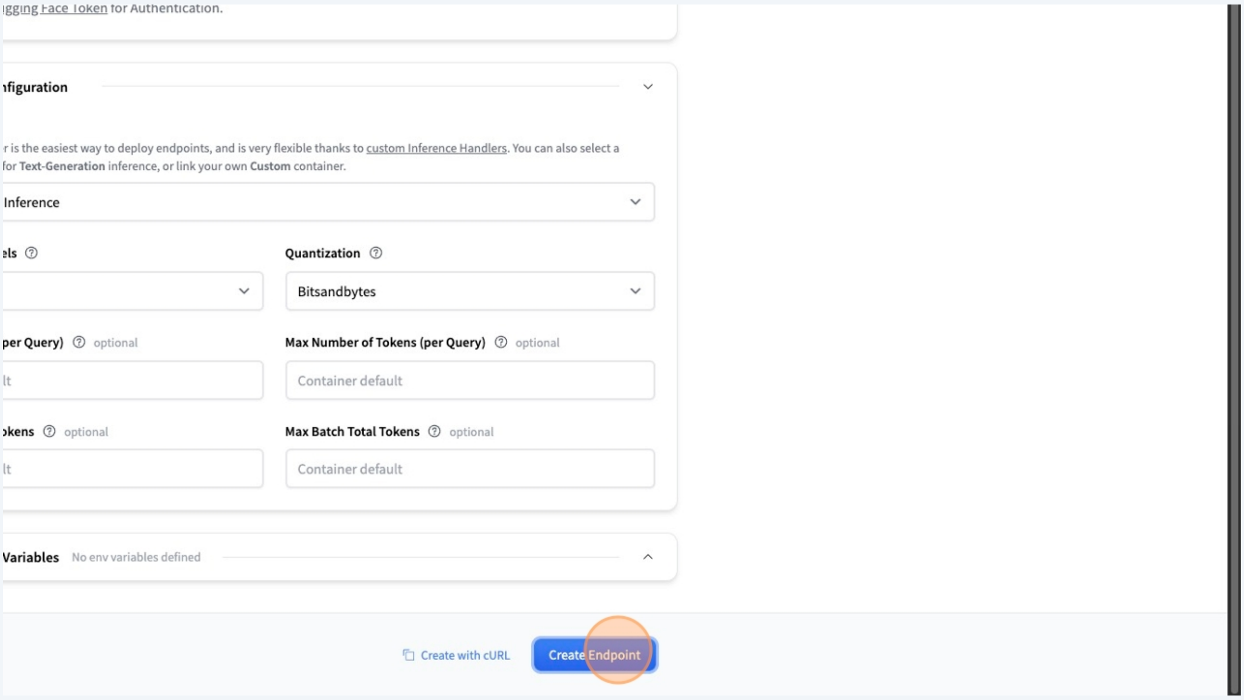
Max Number of Tokens (per Query) optional: Container default

ɿkens optional

Max Batch Total Tokens optional: Container default

Variables No env variables defined

12 Click "Create Endpoint"



ɿgging Face Token for Authentication.

ɿfiguration

r is the easiest way to deploy endpoints, and is very flexible thanks to custom Inference Handlers. You can also select a for Text-Generation inference, or link your own Custom container.

Inference

ɿels

Quantization: Bitsandbytes

per Query) optional

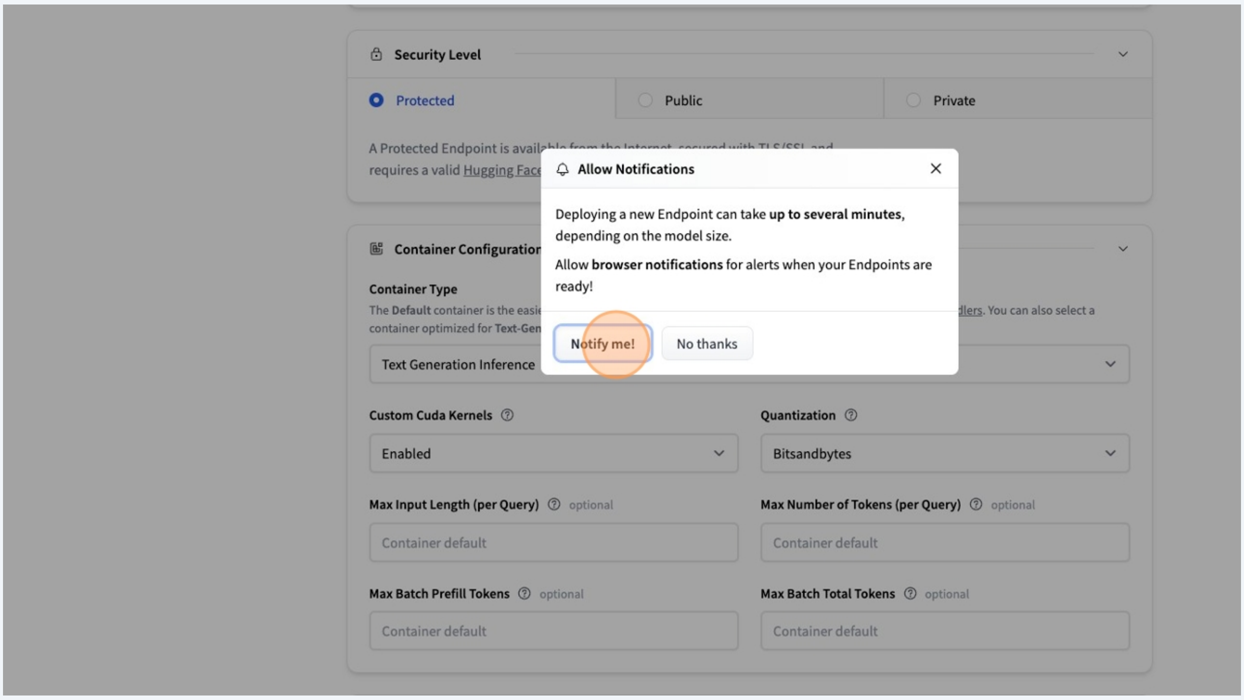
Max Number of Tokens (per Query) optional: Container default

ɿkens optional

Max Batch Total Tokens optional: Container default

Variables No env variables defined

13 Click "Notify me!"



Security Level

● Protected ○ Public ○ Private

A Protected Endpoint is available from the Internet, secured with TLS/SSL and requires a valid Hugging Face

Allow Notifications ×

Deploying a new Endpoint can take **up to several minutes**, depending on the model size.

Allow **browser notifications** for alerts when your Endpoints are ready!

Notify me! No thanks

Container Configuration

Container Type

The **Default** container is the easie ... dlers. You can also select a container optimized for **Text-Gen**

Text Generation Inference

Custom Cuda Kernels: Enabled

Quantization: Bitsandbytes

Max Input Length (per Query) optional: Container default

Max Number of Tokens (per Query) optional: Container default

Max Batch Prefill Tokens optional: Container default

Max Batch Total Tokens optional: Container default