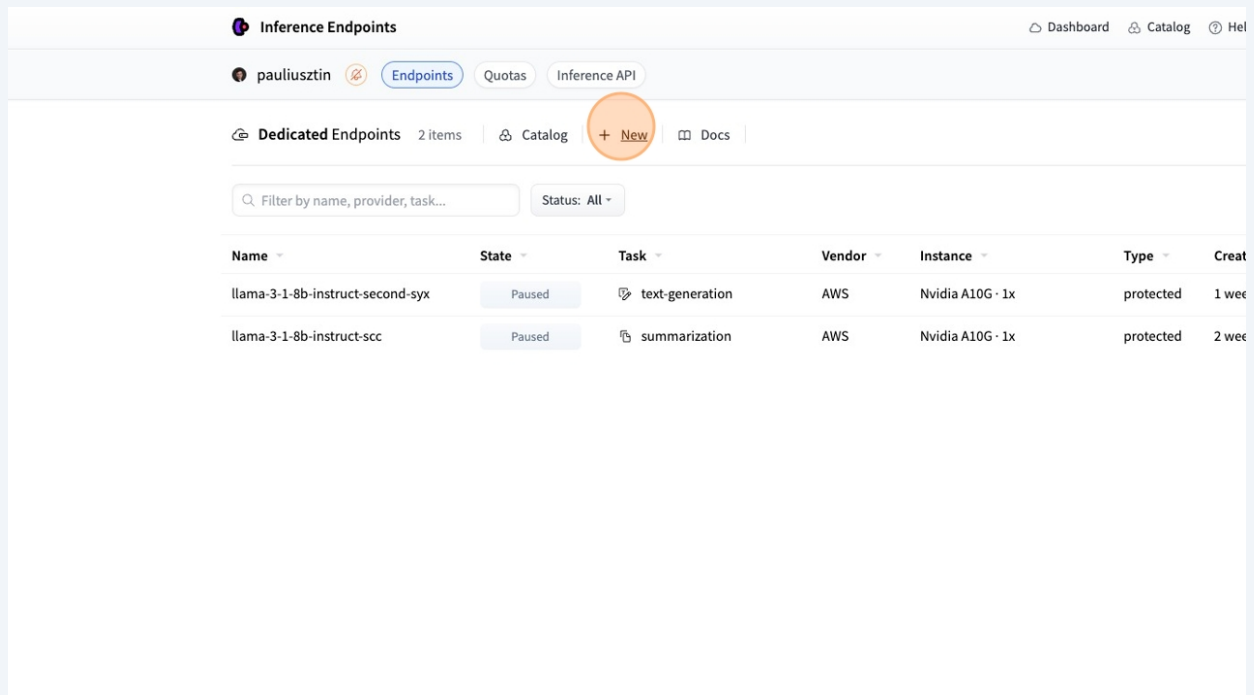


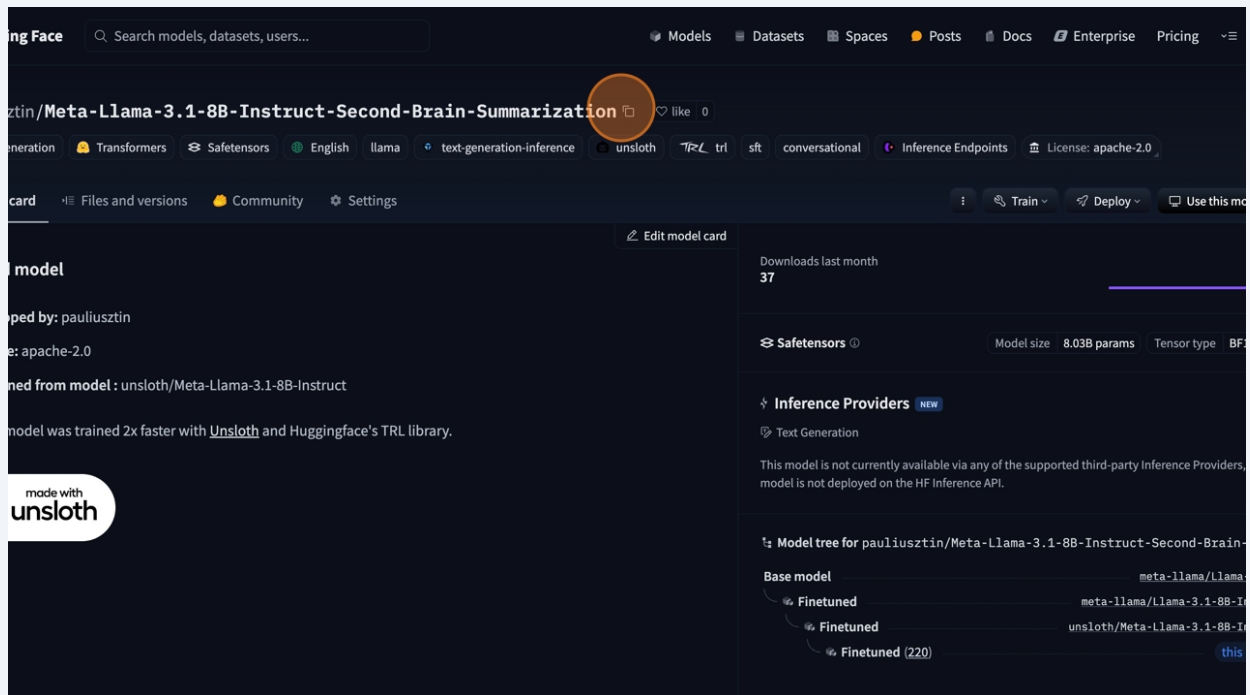
# Creating an Inference Endpoint on Hugging Face

1 Click "New"



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

3 Click this icon.



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.

**Inference Endpoints** Dashboard Cat

[Back to Dashboard](#)

### Create a new Endpoint

[Explore our Inference Catalog](#) to deploy popular models on optimized configuration.

**Model Repository** ? **Endpoint Name** ?

[More options](#)

**Hardware Configuration** ?

[Contact us](#) if you'd like to request a custom solution or instance type.

[aws](#) Amazon Web Services [Microsoft Azure](#) [Google Cloud Platform](#)

**CPU** **GPU** **INF2** N. Virginia us-east-1

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

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

**Inference Endpoints** Dashboard Catalog ? He

[Back to Dashboard](#)

### Create a new Endpoint

[Explore our Inference Catalog](#) to deploy popular models on optimized configuration.

**Model Repository** ? **Endpoint Name** ?

**Models**

**Hardware Configuration** ?

[Contact us](#) if you'd like to request a custom solution or instance type.

[aws](#) Amazon Web Services [Microsoft Azure](#) [Google Cloud Platform](#)

**CPU** **GPU** **INF2** N. Virginia us-east-1

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

## 7 Click "GPU"

Model Repository ⓘ

Endpoint Name ⓘ

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

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 Platform

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 Rapids 8 vCPUs · 16 GB \$ 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 5 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]"

The screenshot shows the AWS SageMaker console's hardware configuration section. At the top, the 'Model Repository' is set to 'pauliusztin/Meta-Llama-3.1-8B-Instruct-Second-B...' and the 'Endpoint Name' is 'meta-llama-3-1-8b-instruct-s-xei'. Below this, the 'Hardware Configuration' section is expanded, showing a link to 'Contact us if you'd like to request a custom solution or instance type.' The 'AWS Amazon Web Services' provider is selected. The 'Region' is set to 'Ireland [eu-west-1]'. The 'Instance Type' is set to 'Nvidia A10G' with '1 GPU · 24 GB' and a price of '\$ 1/h'. Other options include 'Nvidia T4' (1 GPU · 16 GB, \$ 0.5/h), 'Nvidia A100' (1 GPU · 80 GB, \$ 4/h), 'Nvidia A10G' (4 GPUs · 96 GB, \$ 5/h), 'Nvidia A100' (2 GPUs · 160 GB, \$ 8/h), 'Nvidia A100' (4 GPUs · 320 GB, \$ 16/h), and 'Nvidia A100' (8 GPUs · 640 GB, \$ 32/h). The 'Autoscaling' section shows '0 to 1 Replica / Scale-to-zero after 15 min'.

Model Repository <sup>?</sup> pauliusztin/Meta-Llama-3.1-8B-Instruct-Second-B... <sup>?</sup> Endpoint Name meta-llama-3-1-8b-instruct-s-xei

► More options

Hardware Configuration <sup>?</sup>

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

aws Amazon Web Services Microsoft Azure Google Cloud Platform

CPU GPU INF2 Ireland eu-west-1

Instance Type	Configuration	Price
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

The screenshot shows the AWS SageMaker console's configuration section. The 'Instance Type' is set to 'Nvidia A100' with '2 GPUs · 160 GB' and a price of '\$ 8/h'. The 'Autoscaling' section shows '0 to 1 Replica / Scale-to-zero after 15 min'. The 'Visibility' section has 'Public' selected. The 'Configuration' section is set to 'Text Generation Inference'. The 'Variables' section shows 'No env variables defined'. At the bottom, there are buttons for 'Create with cURL' and 'Create Endpoint'.

Instance Type	Configuration	Price
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

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

Public Private

Configuration Text Generation Inference

Variables No env variables defined

Create with cURL Create Endpoint

## 11 Select the "Bitsandbytes" quantization option

int is available from the Internet, secured with TLS/SSL and  
ugging [Face Token](#) for Authentication.

**Configuration** ▾

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** ▾

<b>Models</b> ⓘ	<b>Quantization</b> ⓘ
<input type="text"/>	<input type="text" value="Bitsandbytes"/>
<b>Max Tokens per Query</b> ⓘ optional	<b>Max Number of Tokens (per Query)</b> ⓘ optional
<input type="text"/>	<input type="text" value="Container default"/>
<b>Max Tokens</b> ⓘ optional	<b>Max Batch Total Tokens</b> ⓘ optional
<input type="text"/>	<input type="text" value="Container default"/>

**Variables** No env variables defined ▴

## 12 Click "Create Endpoint"

ugging [Face Token](#) for Authentication.

**Configuration** ▾

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** ▾

<b>Models</b> ⓘ	<b>Quantization</b> ⓘ
<input type="text"/>	<input type="text" value="Bitsandbytes"/>
<b>Max Tokens per Query</b> ⓘ optional	<b>Max Number of Tokens (per Query)</b> ⓘ optional
<input type="text"/>	<input type="text" value="Container default"/>
<b>Max Tokens</b> ⓘ optional	<b>Max Batch Total Tokens</b> ⓘ optional
<input type="text"/>	<input type="text" value="Container default"/>

**Variables** No env variables defined ▴

[Create with cURL](#) [Create Endpoint](#)

### 13 Click "Notify me!"

