

# Building Accelerated AI with Hugging Face + NVIDIA

GTC March 2024

Jeff Boudier, Product @ 





# GPU Poor No More!

Train with H100s,  
no-code,  
serverless,  
on the Hub.

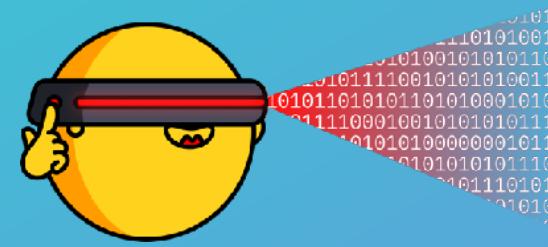


?



[jeff@hf.co](mailto:jeff@hf.co)





6

# things about



you probably didn't know!



1

Hugging Face



# Democratize Good Machine Learning

open source, community, ethics-first

2

# Hugging Face

 **MMitchell** @mmitchell\_ai · Mar 2 ...  
We've been operationalizing ethics [@huggingface](#) along several dimensions, including creating guidelines & **charters** project-by-project. Today, we share our newest guidelines, created for the Diffusers library.  
 Thanks to [@GiadaPistilli](#) for leading!

# ETHICAL GUIDELINES



for developing the **Diffusers library**

[huggingface.co](https://huggingface.co)  
Ethical Guidelines for developing the Diffusers library

[hf.co/ethics](https://hf.co/ethics)

[hf.co/blog/ethics-diffusers](https://hf.co/blog/ethics-diffusers)

 **Hugging Face**  

 **Spaces:**  [society-ethics/about](#)   like 50  Running

  Linked Models  Linked Datasets

 App  Files  Community 5

## Ethics & Society at Hugging Face

At Hugging Face, we are committed to operationalizing ethics at the cutting-edge of machine learning. This page is dedicated to highlighting projects – inside and outside Hugging Face – in order to encourage and support more ethical development and use of AI. We wish to foster ongoing conversations of ethics and values; this means that this page will evolve over time, and your feedback is invaluable. Please open up an issue in the [Community tab](#) to share your thoughts!

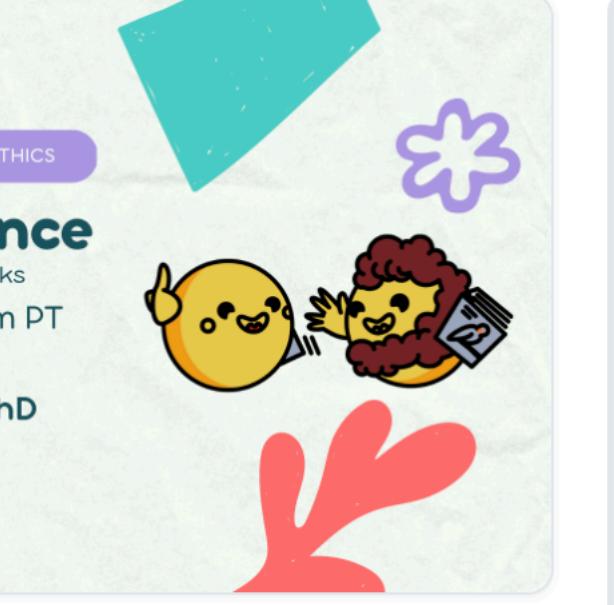
Upcoming Events

 DISCORD Q&A  HUGGINGFACE.CO/ETHICS

**Making Intelligence**  
Ethical Values in IQ and ML Benchmarks

13th Mar. 2023 9:00 am PT

A Q&A with  
Borhane Blili-Hamelin, PhD  
Leif Hancox-Li, PhD



[About the Event](#) [Speaker Bios](#) [Paper Abstract](#)

For our inaugural Ethics & Society Q&A, we're welcoming [Borhane Blili-Hamelin, PhD](#), and [Leif Hancox-Li, PhD](#)!

Come discuss their recent paper (["Making Intelligence: Ethical Values in IQ and ML Benchmarks"](#))

3

Hugging Face

**transformers**

**datasets**

**peft**

**diffusers**

**tokenizers**



20. huggingface

★ 313770

**text-generation-inference**

**optimum**

**trl**

**chat-ui**

**timm**

**nanotron**

**evaluate**

**accelerate**

**setfit**

**safetensors**

**datatrove**

**candle**

**lighteval**

4



# >500k free public models

 **Hugging Face**  Models Datasets Spaces Docs Solutions Pricing ☰ 

Tasks Libraries Datasets Languages Licenses  
Other

Multimodal

Feature Extraction Text-to-Image  
Image-to-Text Image-to-Video  
Text-to-Video Visual Question Answering  
Document Question Answering  
Graph Machine Learning Text-to-3D  
Image-to-3D

**Models 431,018** Filter by name new Full-text search ↑↓ Sort: Trending

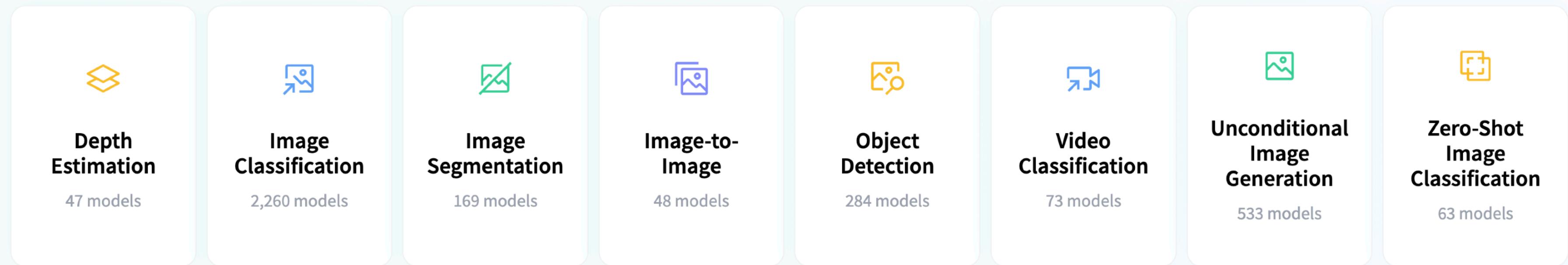
- s. stabilityai/sdxl-turbo**  
Text-to-Image • Updated 4 days ago • ↓ 443k • ❤ 1.16k
- playgroundai/playground-v2-1024px-aesthetic**  
Text-to-Image • Updated 3 days ago • ↓ 118k • ❤ 292
- mistralai/Mixtral-8x7B-Instruct-v0.1**  
Text Generation • Updated about 11 hours ago • ↓ 4.51k • ❤ 273
- DiscoResearch/mixtral-7b-8expert**  
Text Generation • Updated about 16 hours ago • ↓ 8.86k • ❤ 218
- mistralai/Mixtral-8x7B-v0.1**

# /tasks for all ML use cases

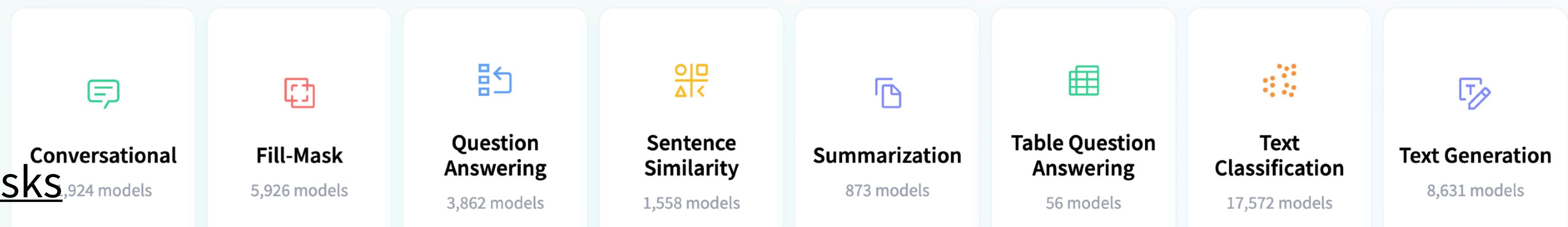
## Tasks

Hugging Face is the home for all Machine Learning tasks. Here you can find what you need to get started with a task: demos, use cases, models, datasets, and more!

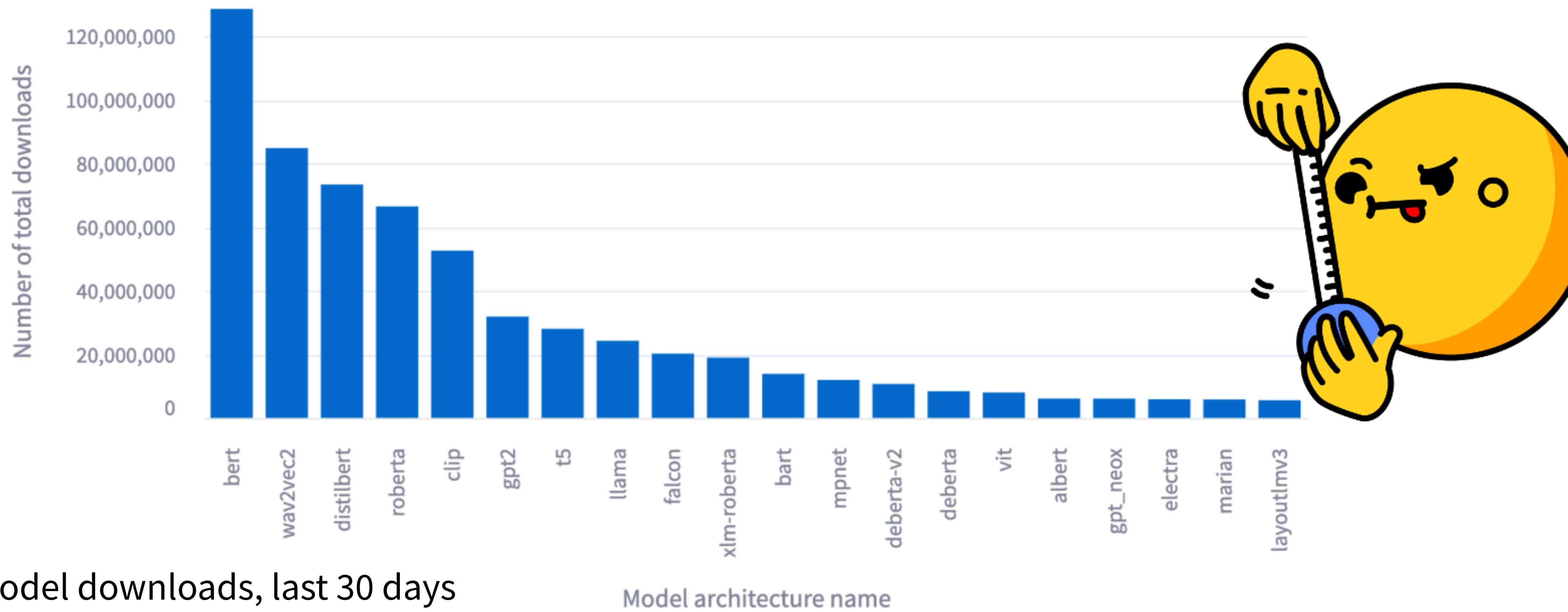
### Computer Vision



### Natural Language Processing



# >10M model downloads /day



Transformers model downloads, last 30 days

Source: [hf.co/spaces/huggingface/transformers-stats](https://hf.co/spaces/huggingface/transformers-stats)



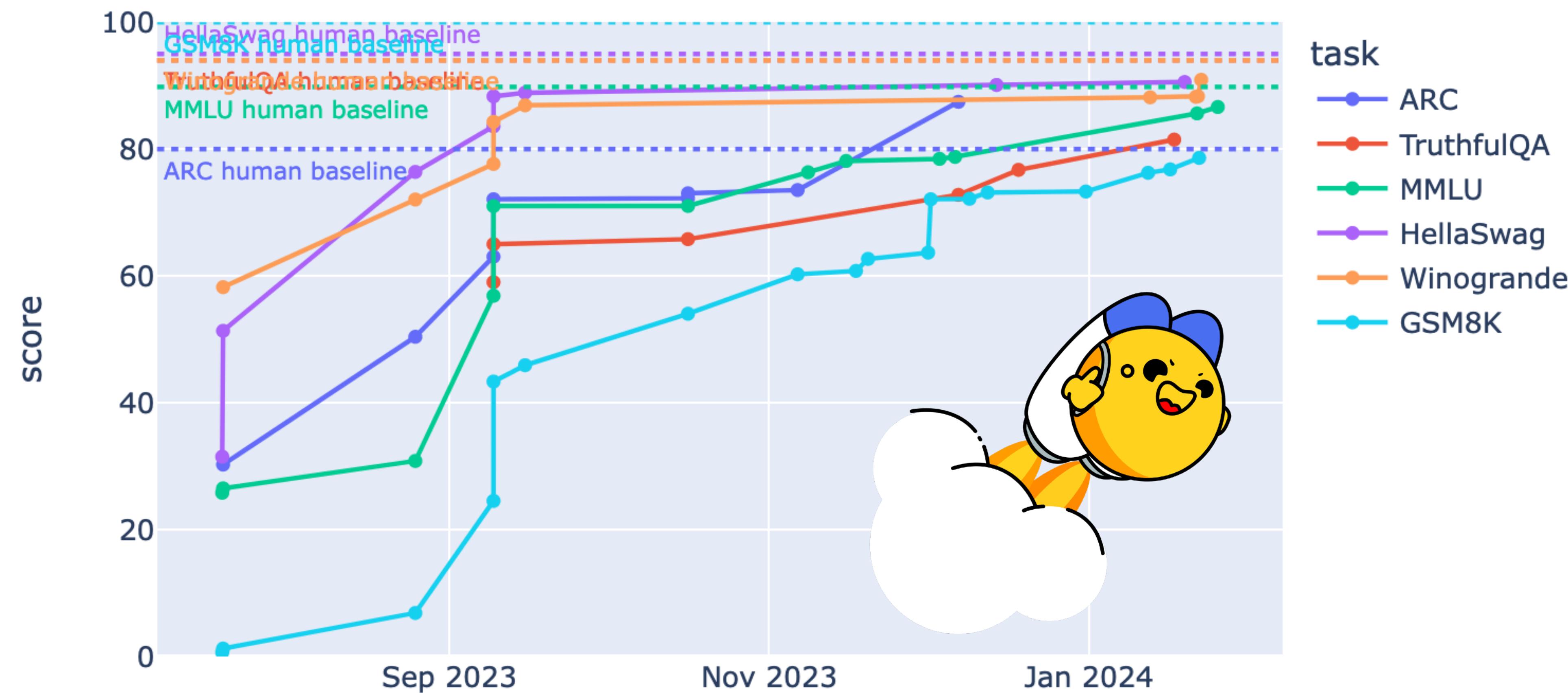
# Why you should build AI with open source

1

# Future-proof: open models



Top Scores and Human Baseline Over Time (from last update)



2

# Control costs: open ML efficiency



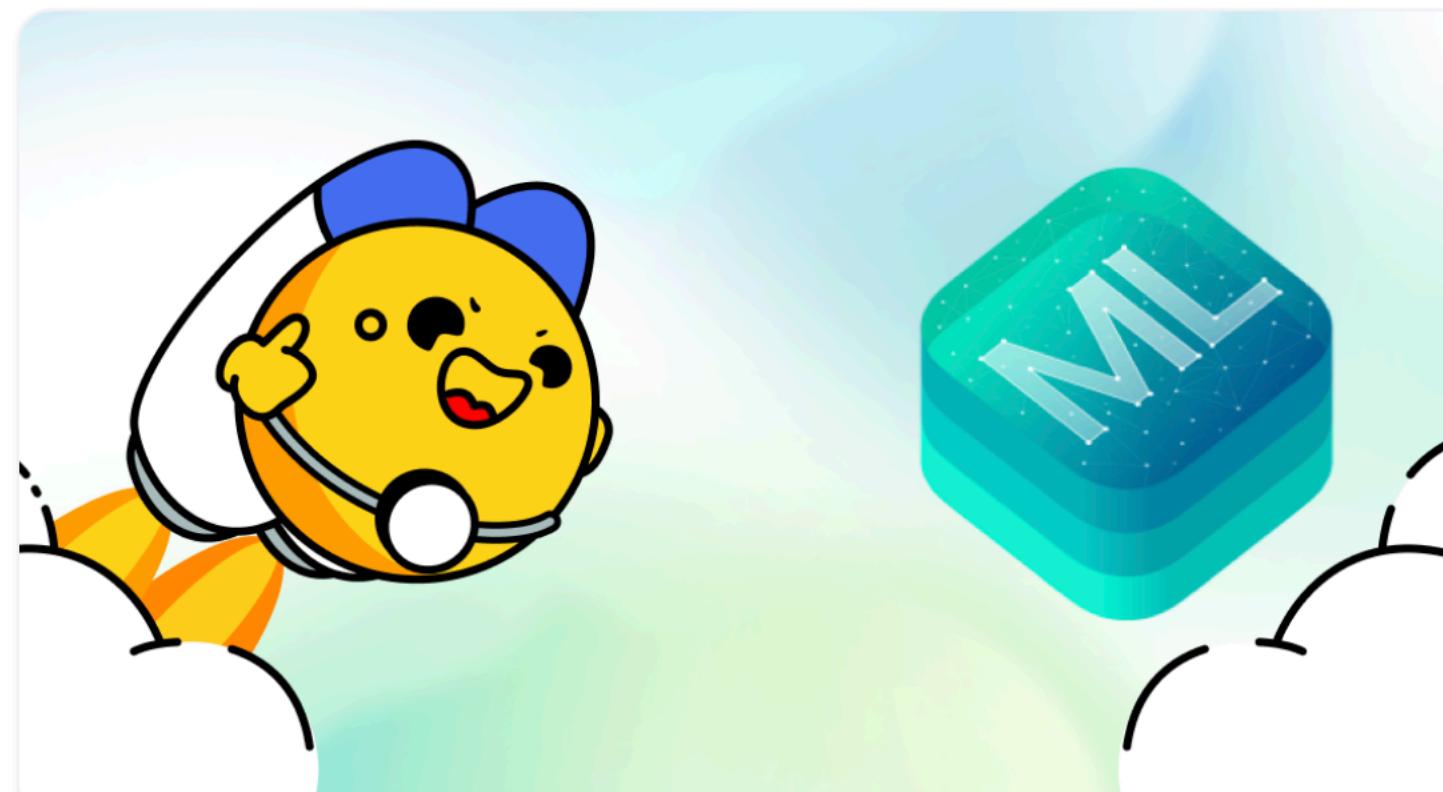
**Exploring simple optimizations for SDXL**

By sayakpaul • October 24, 2023



**Accelerating Stable Diffusion XL Inference with JAX on Cloud TPU v5e**

By pcuenq • October 3, 2023



**Faster Stable Diffusion with Core ML on iPhone, iPad, and Mac**

By pcuenq • June 15, 2023

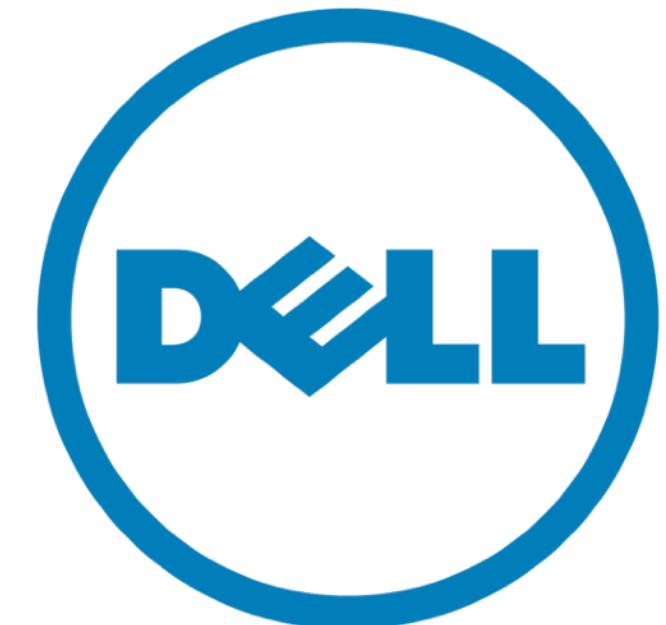
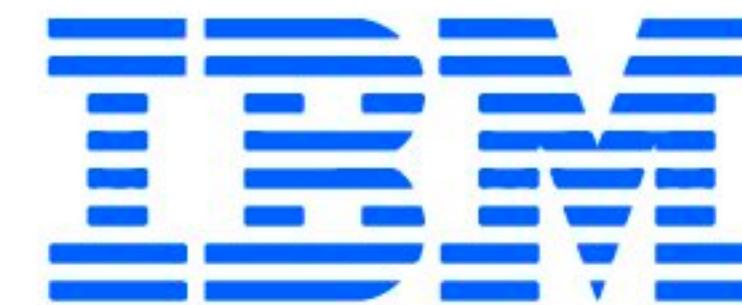


**SDXL in 4 steps with Latent Consistency LoRAs**

By pcuenq • November 9, 2023

3

# Security: Your model, your hosting



4

# Models you (version) control



s. stabilityai/sdxl-turbo like 1.16k

Text-to-Image Diffusers ONNX Safetensors StableDiffusionXLPipeline License: sai-nc-community (other)

Model card Files Community 25

main sdxl-turbo 11 contributors History: 35 commits + Contribute

Commit History

update the readme to fix the image-to-image example (#18) f4b0486   
patrickvonplaten HF STAFF vikasp committed on 4 days ago

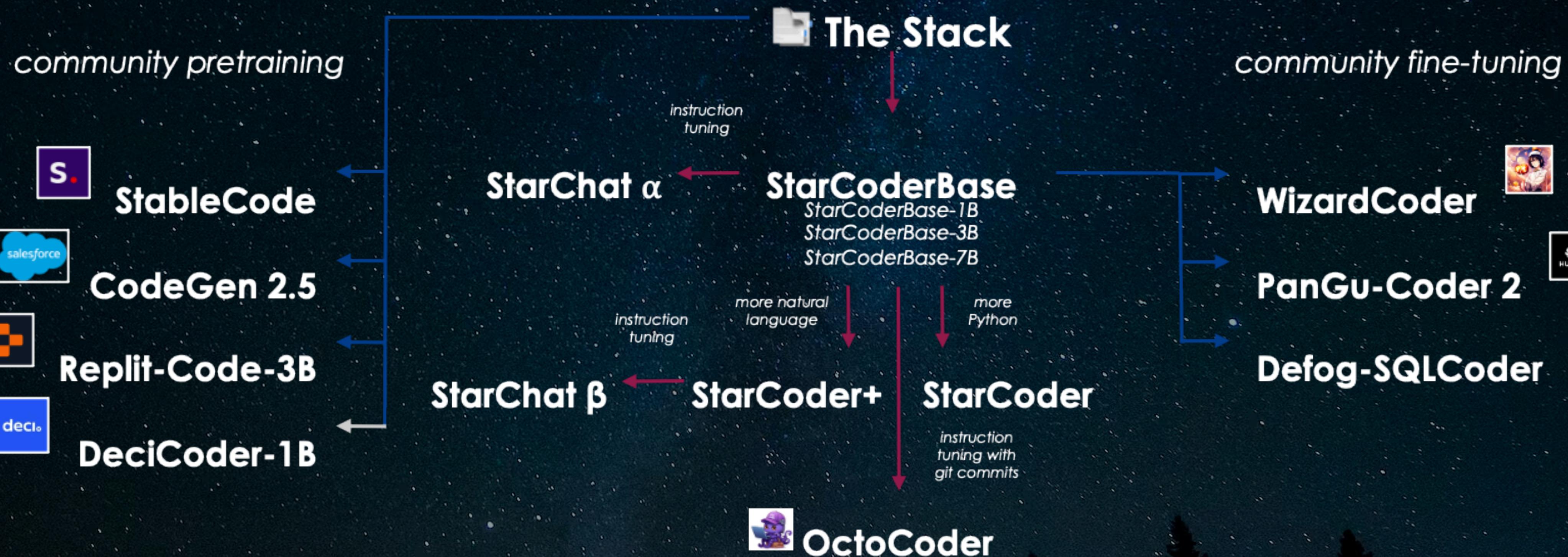
Update README.md (#19) 1beb956   
patrickvonplaten HF STAFF matospiso committed on 4 days ago

onnx (#15) fbda352   
patrickvonplaten HF STAFF echarlaix HF STAFF committed on 8 days ago

# 5 Models you can trust (and verify)

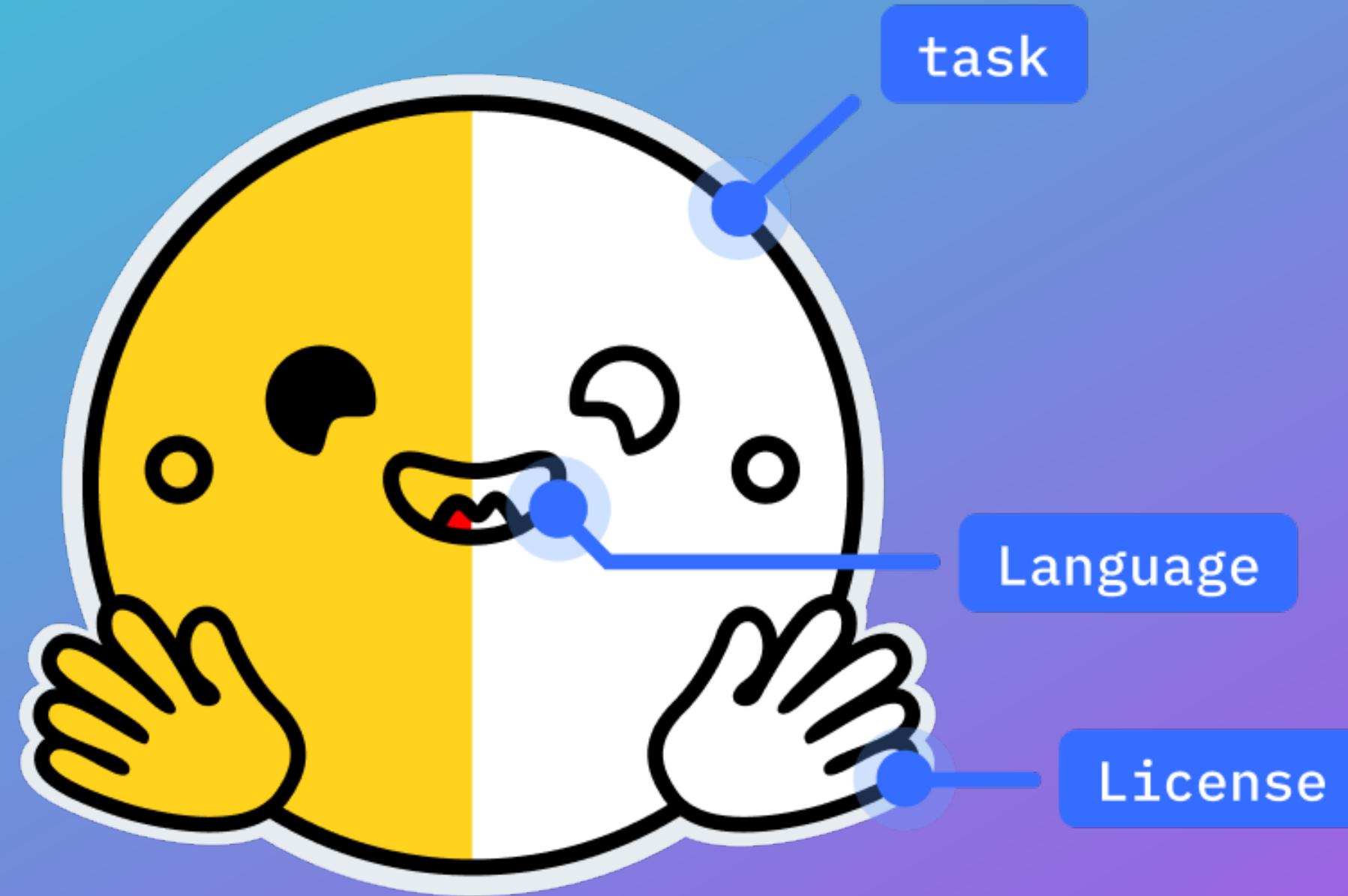


## BigCode Ecosystem



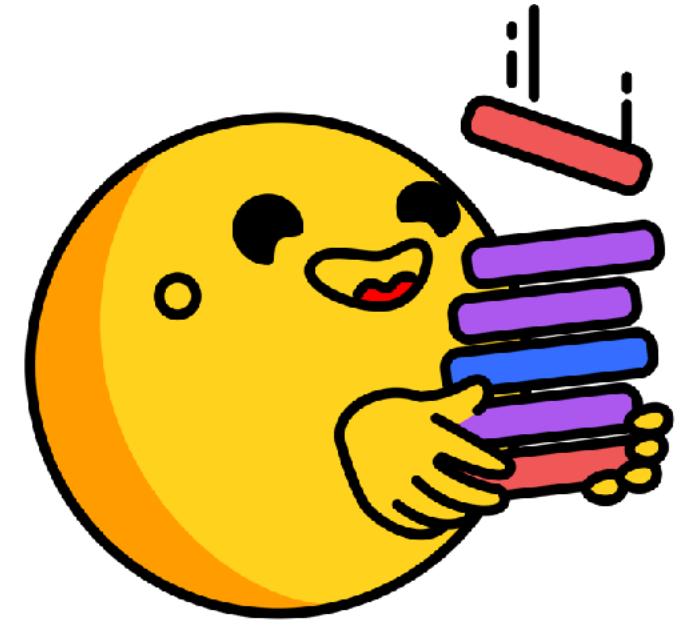
# Why build AI with open source?

- 1 Future-proof: open models 
- 2 Control costs: open ML efficiency 
- 3 Security: Your model, your hosting 
- 4 Models you (version) control 
- 5 Models you can trust (and verify) 



# How to build AI with 😊 and open source

Building Accelerated AI with Hugging Face + NVIDIA - GTC March 2024



# 60,000+ open LLMs

<b>Grok-1</b>	<b>Command-R</b>		<b>StarCoder 2</b>
<b>CodeLLaMa</b>	<b>Gemma</b>	<b>Mixtral</b>	<b>Phi-2</b>
<b>LLaMa 2</b>	<b>FALCON</b>	<b>Mistral</b>	<b>Zephyr</b>
<b>LLaMa</b>	<b>Alpaca</b>	<b>BLOOM</b>	<b>RedPajama-INCIT</b>
	<b>StableLM-Base</b>		
<b>ChatGLM</b>	<b>Galactica</b>	<b>Dolly</b>	<b>MPT</b>
<b>Pythia</b>	<b>BLOOMZ</b>	<b>GPT-JT</b>	<b>CodeCapybara</b>
		<b>Vicuna</b>	<b>Koala</b>
<b>Open-Assistant SFT</b>	<b>T5</b>		<b>StackLlama SFT</b>
<b>StackLlama RLHF</b>	<b>Flan-UL2</b>	<b>StableLM-Tuned</b>	
		<b>Flan-T5</b>	<b>Palmyra-Base</b>
	<b>Camel</b>		<b>vicuna-13b-fine-tuned-rlhf</b>

[Click here to log in through Single Sign-On to view activity within the huggingface org.](#)

# LLMs: Compare them all!

Open LLM Leaderboard

The Open LLM Leaderboard aims to track, rank, and evaluate open LLMs across benchmarks. Submit a model for automated evaluation on the GPU cluster on the "Submit" page!

The leaderboard's backend runs the great [Eleuther AI Language Model Evaluation Harness](#) to compute numbers. Read more details and reproducibility on the "About" page!

Other cool benchmarks for LLMs are developed at HuggingFace: [human and GPT4 evals](#), [performance benchmarks](#)

And also in other labs, check out the [AlpacaEval Leaderboard](#) and [MT Bench](#) among other great resources!

**LLM Benchmark**   [About](#)   [Submit here!](#)

Select columns to show

Average    ARC    HellaSwag    MMLU    TruthfulQA  
 Type    Precision    Hub License    #Params (B)    Hub ❤️  
 Model sha

Search for your model and press ENTER  
 Filter model types  
 all    pretrained    fine-tuned    instruction  
 RL-tuned

T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA
●	<a href="#">meta-llama/Llama-2-70b-hf</a>	67.35	67.32	87.33	69.83	44.92
●	<a href="#">huggyllama/llama-65b</a>	64.23	63.48	86.09	63.93	43.43
●	<a href="#">llama-65b</a>	64.23	63.48	86.09	63.93	43.43
●	<a href="#">lmsys/vicuna-13b-v1.5</a>	61.69	57	81.23	56.87	51.67
●	<a href="#">llama-30b</a>	61.68	61.26	84.73	58.47	42.27

# 100,000 organizations

| Enterprise Hub

**Enterprise-ready version of  
the world's leading AI platform**

Subscribe to  Enterprise Hub

for \$20/user/month with your Hub organization

Give your organization the most advanced platform to build AI with  
enterprise-grade security, access controls, dedicated support and more.

# NVIDIA #1

# Enterprise Hub organization

The screenshot shows the GitHub organization profile for NVIDIA. At the top, there's a green square icon with a white stylized eye logo, followed by the text "NVIDIA" in bold black, "Enterprise" in a smaller gray font, and "Company" in a purple font. Below this is a URL "https://www.nvidia.com/" and a GitHub handle "nvidia". To the right are "Watch repos" and "i" buttons.

**AI & ML interests**  
None defined yet.

**Team members** 728  
A grid of 728 small circular profile pictures of team members.

**Collections** 7

- Nemotron 3 8B** >  
The Nemotron 3 8B Family of models is optimized for building ...
  - nvidia/nemotron-3-8b-base-4k**  
Text Generation • Updated 29 days ago • ↓ 14 • ❤ 44
  - nvidia/nemotron-3-8b-chat-4k-sft**  
Text Generation • Updated 29 days ago • ↓ 15 • ❤ 4
  - nvidia/nemotron-3-8b-chat-4k-rlhf**  
Text Generation • Updated 29 days ago • ↓ 46 • ❤ 20
- SteerLM** >  
A collection of models and datasets relating to SteerLM and He...
  - nvidia/HelpSteer**  
Viewer • Updated Jan 3 • ↓ 1.79k • ❤ 146
  - nvidia/Llama2-70B-SteerLM-Chat**  
Text Generation • Updated Jan 3 • ❤ 21
  - nvidia/Llama2-13B-SteerLM-RM**  
Text Generation • Updated 15 days ago • ↓ 16 • ❤ 1

# Train it easy with AutoTrain

The screenshot shows the AutoTrain web interface with a large central input form. At the top, the navigation bar includes 'Spaces' (private), 'Running' (highlighted in green), 'App' (selected), 'Files', 'Community', 'Settings', and a user profile icon.

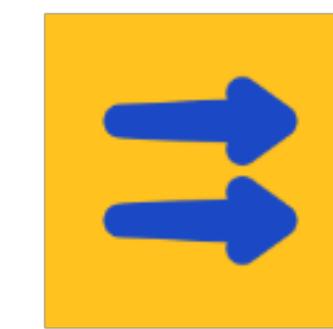
The main form has a title 'auto TRAIN' and a descriptive text box:

AutoTrain Advanced is a no-code solution that allows you to train machine learning models in just a few clicks. Please note that you must upload data in correct format for project to be created. For help regarding proper data format and pricing, click [here](#).

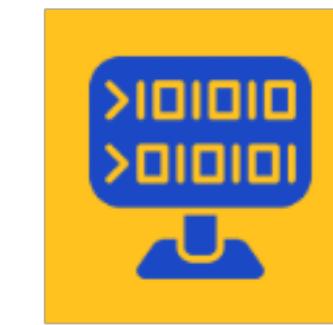
The form fields are as follows:

- Hugging Face User:** jeffboudier
- Project name:** nvidia-gtc-demo
- Task:** LLM SFT
- Base Model:** gface.co/google/gemma-7b
- Hardware:** A10G Large
- Training Data:** (with a cloud-upload icon)
- Upload Training File(s):**

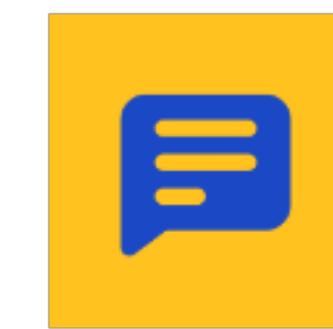
# text-generation-inference



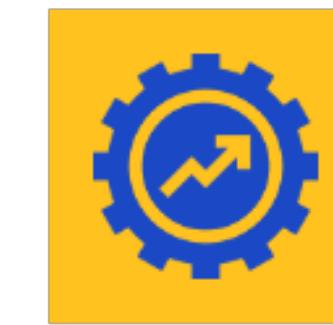
Tensor  
Parallelism



Quantization



Token Streaming



Optimizations



Metrics and  
monitoring



Security

Production solution for max throughput

Gemma - Mistral - Llama v2 - Falcon - StarCoder 2 - GPT-NeoX

Optimized for NVIDIA GPU, AMD GPU, AWS Inferentia2, Habana Gaudi

# HF Inference Endpoints



## Model Catalog

Browse our selection of hand-picked, ready-to-deploy models!

Filter by name, author...

Task: All ▾

All Hub Models

### Text Generation

#### Mixtral-8x7B-Instruct-v0.1

TGI

Text Generation • mistralai

Mixtral 8x7B is a sparse mixture-of-experts decoder-only model fine-tuned on instruction following a permissive license.

GPU 2x Nvidia A100 ▾

\$ 13 / h

Go

#### Llama-2-70B-chat-GPTQ

TGI

Text Generation • TheBloke

70-billion parameters model from Meta, optimized for dialogue. Generates helpful, safe responses and outperforms other open-source chat LLMs.

GPU 2x Nvidia A100 ▾

\$ 13 / h

Go

#### Llama-2-13B-chat-GPTQ

TGI

Text Generation • TheBloke

13-billion parameters model from Meta, optimized for dialogue. Generates helpful, safe responses and outperforms other open-source chat models.

GPU 1x Nvidia A10G ▾

\$ 1.3 / h

Go

#### Falcon-180B-Chat-GPTQ

TGI

Text Generation • TheBloke

180-billion parameters conversational AI model from TII, optimized for fast inference through an efficient architecture. Freely available under TII LICENSE.

GPU 2x Nvidia A100 ▾

\$ 13 / h

Go

#### Mistral-7B-Instruct-v0.1

TGI

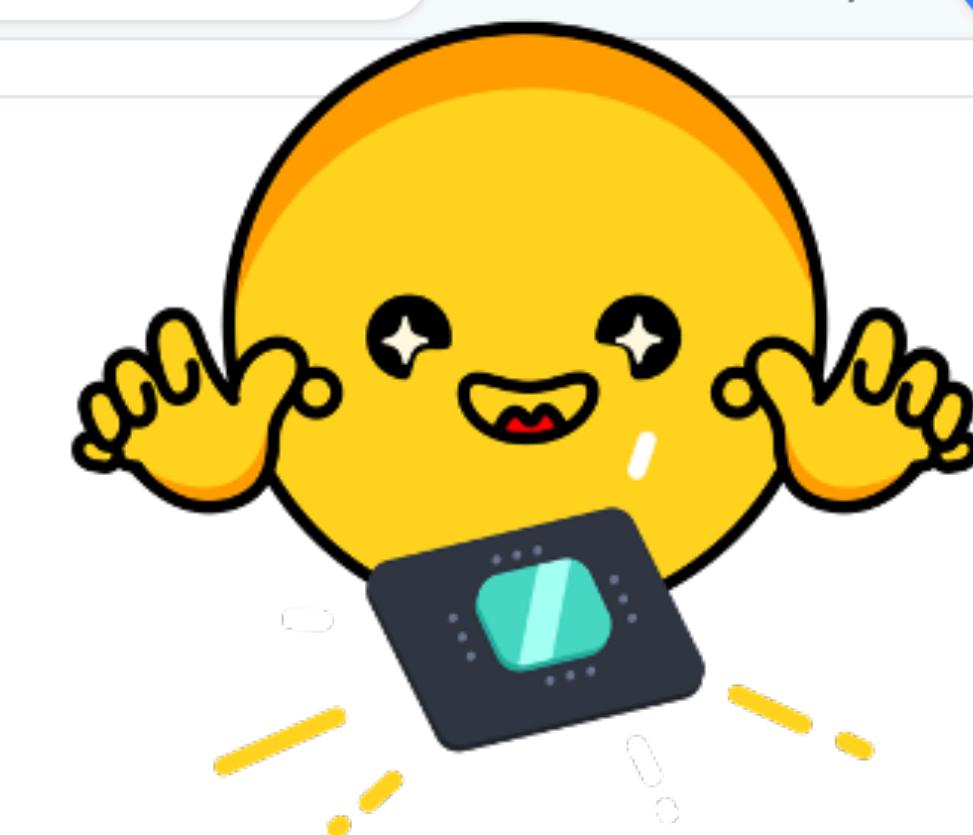
Text Generation • mistralai

7-billion parameters model from Mistral AI, fine-tuned using a variety of publicly available conversation datasets.

GPU 1x Nvidia A10G ▾

\$ 1.3 / h

Go



# Ready, Set... Demo



Discover amazing ML apps made by the community!

Create new Space

or [learn more about Spaces.](#)

Search Spaces

new Full-text search

↑↓ Sort: Trending

## ☆ Spaces of the week 🔥

Running on **A10G** ❤ 80

**Playground V2**

playgroundai 6 days ago

Running on **A100** ❤ 103

**Enhance This Demofusion S...**

radames about 3 hours ago

AI Tube ❤ 29

**AI Tube**

jbilcke-hf about 16 hours ago

Running on **A10G** ❤ 65

**Marigold Depth Estimation**

toshas 4 days ago

Running on **CPU UPGRADE** ❤ 26

**NexusRaven-V2 Demo**

Nexusflow 6 days ago

Running on **A10G** ❤ 77

**Seine**

Vchitect 5 days ago

Tryemoji ❤ 10

**Tryemoji**

leptonai 12 days ago

Like History ❤ 33

**Like History**

timqian 5 days ago

# 500,000 Spaces created

Spaces | nvidia/canary-1b | like 140 | Running on T4

App Files Community 1

## NeMo Canary model: Transcribe & Translate audio

**Step 1:** Upload an audio file or record with your microphone.

This demo supports audio files up to 10 mins long. You can transcribe longer files locally with this NeMo [script](#).

**Step 2:** Choose the input and output language.

**Step 3:** Run the model.

**Run model**

**Model Output**

Quelle est la température de l'eau glaciale à Fahrenheit ?

Audio

0:00 0:06

1x

1x

Up Microphone

Input audio is spoken in:

English

Transcribe in language:

# Accelerate Spaces with NVIDIA

## Space Hardware (i)

Choose a hardware for your Space.

You'll be billed on a per minute basis.

View usage in your [billing settings](#).

Display price:  per hour  per month

### CPU basic

2 vCPU · 16 GB RAM

Free

### Zero Nvidia A100

Dynamic resources

Free

### CPU upgrade

8 vCPU · 32 GB RAM

\$0.03/hour

### CPU XL

16 vCPU · 124 GB RAM

Free

### Nvidia T4 small

4 vCPU · 15 GB RAM · 16 GB VRAM

\$0.60/hour

### Nvidia T4 medium

8 vCPU · 30 GB RAM · 16 GB VRAM

\$0.90/hour

### Nvidia A10G small

4 vCPU · 15 GB RAM · 24 GB VRAM

\$1.05/hour

### Nvidia A10G large

12 vCPU · 46 GB RAM · 24 GB VRAM

\$3.15/hour

### Nvidia A100 large

12 vCPU · 142 GB RAM · 40 GB VRAM

\$4.13/hour

### Nvidia 2xA10G large

24 vCPU · 92 GB RAM · 48 GB VRAM

\$5.70/hour

### Nvidia 4xA10G large

48 vCPU · 184 GB RAM · 96 GB VRAM

\$10.80/hour

### Nvidia H100 large

24 vCPU · 250 GB RAM · 80 GB VRAM

\$8.70/hour

## Sleep time settings (i)

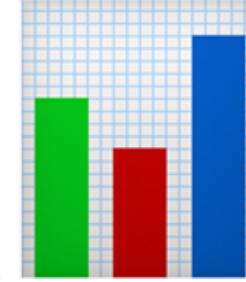
Sleep after  of inactivity

Upgrade to a paid Hardware to set a custom sleep time.

### AI Accelerator

HPU · IPU · ...

# How to build AI with Hugging Face



**Assess models with Leaderboards**



**Collaborate with Enterprise Hub**



**Train with AutoTrain**



**Deploy with Inference Endpoints**



**Demo with Spaces**



# You're not alone

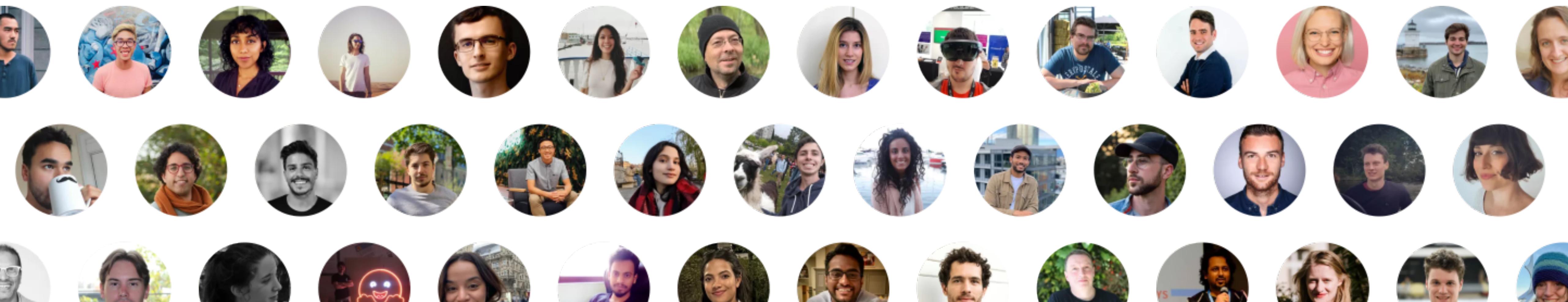
## Build your own ML, faster

Build better AI features in-house with open source and Hugging Face experts by your side to guide you along the way.



[Request a Quote](#)

to accelerate your ML roadmap





# Accelerating AI with Hugging Face + NVIDIA

Building Accelerated AI with Hugging Face + NVIDIA - GTC March 2024

Introducing...

Train on DGX Cloud

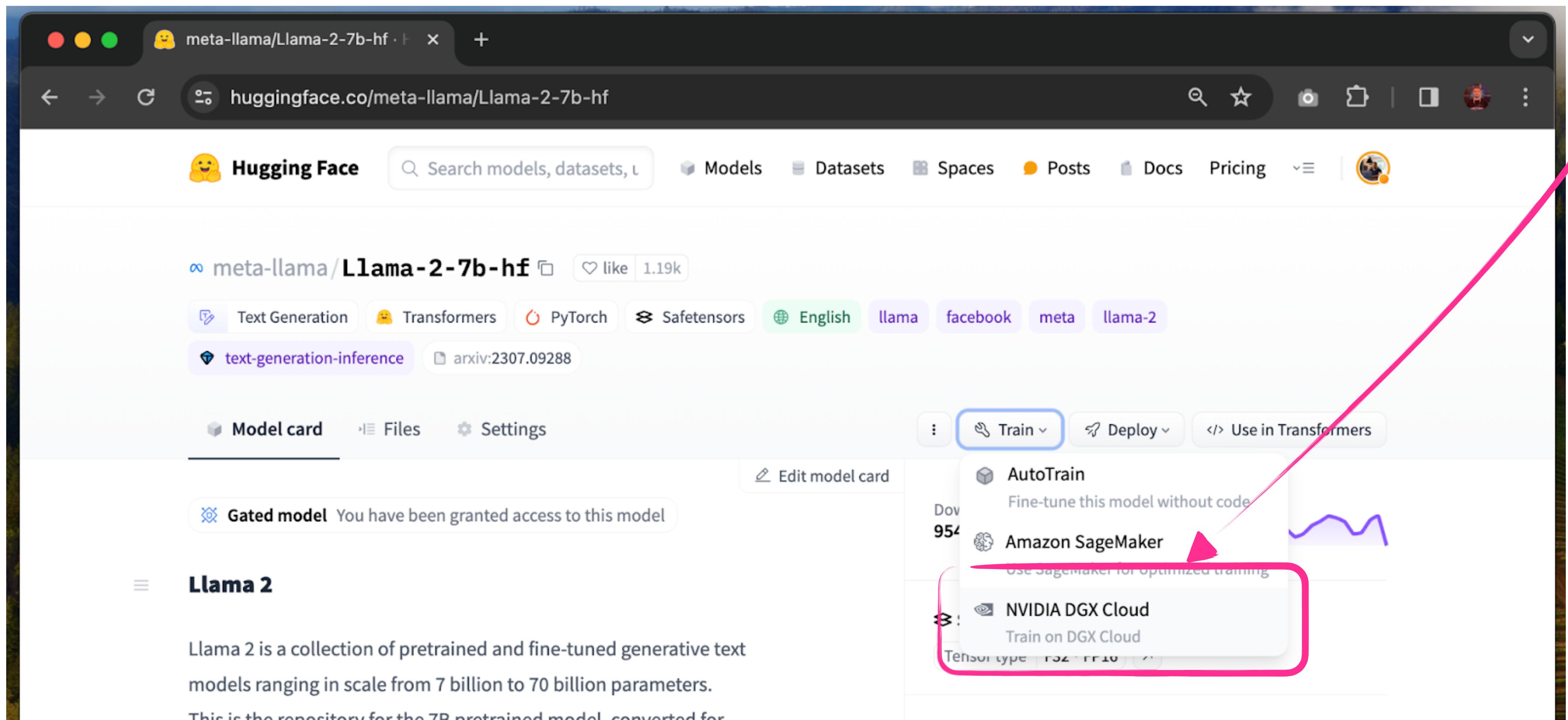




**Finetune LLMs  
on the Hub  
with H100 and L40S  
without code**

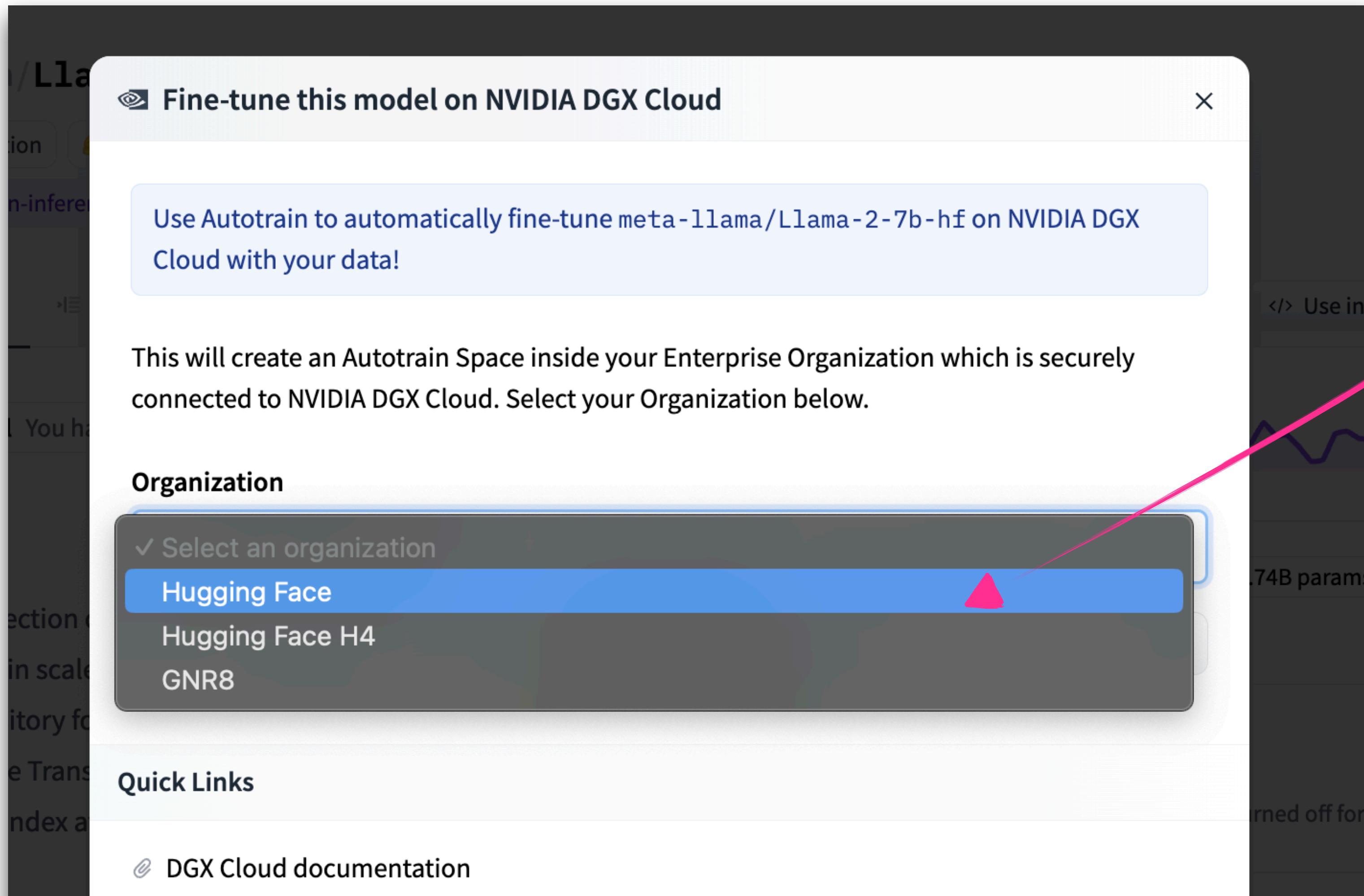
# How it works

## 1. Pick a model, click Train on DGX Cloud



# How it works

## 2. Select your Enterprise Hub organization



# How it works

## 3. Select Hardware, set parameters

The screenshot shows the Autotrain interface for configuring a project. The interface includes fields for Hugging Face User (philschmid), Project name (autotrain-a74ce-65ini), and Training Parameters (a JSON object). A pink arrow points from the 'Hardware' dropdown (1xL40) to the 'Task' dropdown (LLM SFT). Another pink arrow points from the 'Task' dropdown to the 'Training Parameters' section.

**Hugging Face User**: philschmid

**Project name**: autotrain-a74ce-65ini

**Training Parameters** (find params to copy-paste [here](#)):

```
{  
    "block_size": 1024,  
    "model_max_length": 2048,  
    "padding": "right",  
    "use_flash_attention_2": false,  
    "disable_gradient_checkpointing": false,  
    "logging_steps": -1,  
    "evaluation_strategy": "epoch",  
    "save_total_limit": 1,  
    "save_strategy": "epoch",  
    "auto_find_batch_size": false,  
    "mixed_precision": "fp16",  
    "lr": 0.00003,  
    "epochs": 3,  
    "batch_size": 2,  
    "warmup_ratio": 0.1,  
    "gradient_accumulation": 1,  
    "optimizer": "adamw_torch",  
    "scheduler": "linear",  
    "weight_decay": 0,  
    "max_grad_norm": 1,  
    "seed": 42,  
    "chat_template": "none",  
}
```

**Hardware**: 1xL40

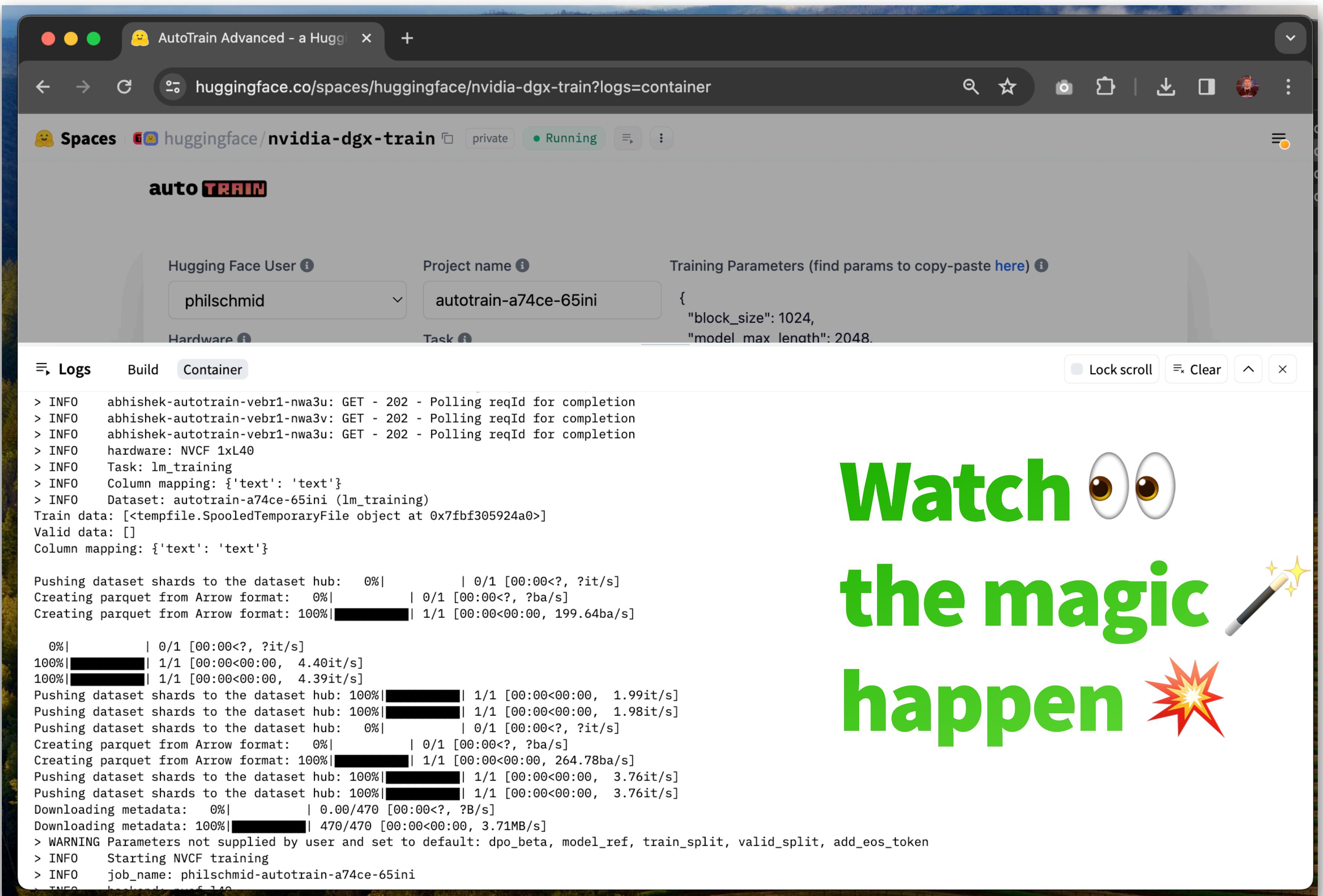
**Task**: LLM SFT

**Base Model**: mistralai/Mistral-7B-v0.1

**Training Data**:

**Column mapping**: {"text": "text"}

# How it works

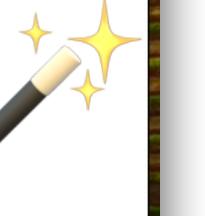


The screenshot shows the AutoTrain Advanced interface in a web browser. The title bar says "AutoTrain Advanced - a Huggi...". The address bar shows "huggingface.co/spaces/huggingface/nvidia-dgx-train?logs=container". The main area has a header "auto TRAIN" with sections for "Hugging Face User" (philschmid), "Project name" (autotrain-a74ce-65ini), and "Training Parameters" (block\_size: 1024, model\_max\_length: 2048). Below this, there are tabs for "Logs", "Build", and "Container", with "Logs" selected. The log output shows the progress of a training job, including dataset loading, parquet creation, and dataset pushing to the hub. The logs end with a warning about missing parameters and starting NVCF training.

```
> INFO abhishek-autotrain-vebr1-nwa3u: GET - 202 - Polling reqId for completion
> INFO abhishek-autotrain-vebr1-nwa3v: GET - 202 - Polling reqId for completion
> INFO abhishek-autotrain-vebr1-nwa3u: GET - 202 - Polling reqId for completion
> INFO hardware: NVCF 1xL40
> INFO Task: lm_training
> INFO Column mapping: {'text': 'text'}
> INFO Dataset: autotrain-a74ce-65ini (lm_training)
Train data: [<tempfile.SpooledTemporaryFile object at 0x7fbf305924a0>]
Valid data: []
Column mapping: {'text': 'text'}

Pushing dataset shards to the dataset hub: 0% | 0/1 [00:00<?, ?it/s]
Creating parquet from Arrow format: 0% | 0/1 [00:00<?, ?ba/s]
Creating parquet from Arrow format: 100%|██████████| 1/1 [00:00<00:00, 199.64ba/s]

0% | 0/1 [00:00<?, ?it/s]
100%|██████████| 1/1 [00:00<00:00, 4.40it/s]
100%|██████████| 1/1 [00:00<00:00, 4.39it/s]
Pushing dataset shards to the dataset hub: 100%|██████████| 1/1 [00:00<00:00, 1.99it/s]
Pushing dataset shards to the dataset hub: 100%|██████████| 1/1 [00:00<00:00, 1.98it/s]
Pushing dataset shards to the dataset hub: 0% | 0/1 [00:00<?, ?it/s]
Creating parquet from Arrow format: 0% | 0/1 [00:00<?, ?ba/s]
Creating parquet from Arrow format: 100%|██████████| 1/1 [00:00<00:00, 264.78ba/s]
Pushing dataset shards to the dataset hub: 100%|██████████| 1/1 [00:00<00:00, 3.76it/s]
Pushing dataset shards to the dataset hub: 100%|██████████| 1/1 [00:00<00:00, 3.76it/s]
Downloading metadata: 0% | 0.00/470 [00:00<?, ?B/s]
Downloading metadata: 100%|██████████| 470/470 [00:00<00:00, 3.71MB/s]
> WARNING Parameters not supplied by user and set to default: dpo_beta, model_ref, train_split, valid_split, add_eos_token
> INFO Starting NVCF training
> INFO job_name: philschmid-autotrain-a74ce-65ini
> INFO hardware: NVCF 1xL40
```

Watch   
the magic   
happen 

# How it works

philschmid/autotrain-snlab-fxf0s private Model card Files and versions Training metrics Community Settings

TensorBoard TIME SERIES SCALARS TEXT INACTIVE

Filter tags (regex) All Scalars Image Histogram Settings

Pinned

## Watch 🐚 the magic 🌟 happen 💥

train 9 cards

train/epoch

Run ↑	Value	Step	Relative
runs/Mar14_09- 53-05_np-atl3-br2-054-b	3	342	7.047 min

train/grad\_norm

Run ↑	Value	Step	Relative
runs/Mar14_09- 53-05_np-atl3-br2-054-b	2,3521	306	6.176 min

train/learning\_rate

Run ↑	Value	Step	Relative
runs/Mar14_09- 53-05_np-atl3-br2-054-b	0	306	6.176 min

train/loss

Run ↑	Value	Step	Relative
runs/Mar14_09- 53-05_np-atl3-br2-054-b	0.9414	306	6.176 min

Settings

GENERAL

Horizontal Axis Step

Enable step selection and data table (Scalars only)

Enable Range Selection

Link by step 342

Card Width

SCALARS

Smoothing

Tooltip sorting method Alphabetical

Ignore outliers in chart scaling

# How it works

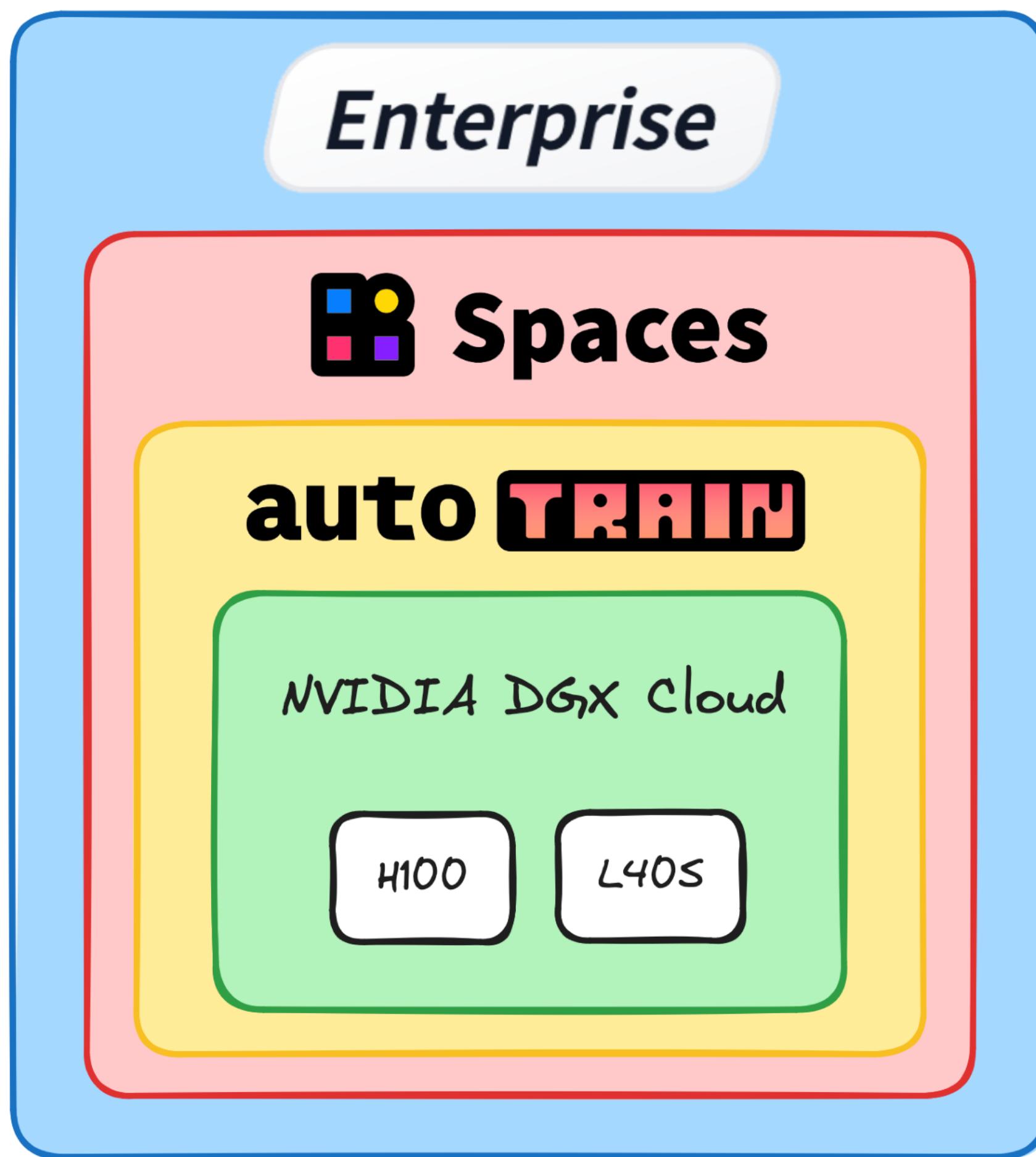
## 4. Check out your new private model!

The screenshot shows a web browser displaying a model card for a private AutoTrain model. The top navigation bar includes the repository name "philschmid/autotrain-snlab-fxf0s" and a "private" status indicator. Below the bar are several tabs: "Text Generation", "TensorBoard", "Safetensors", "Trained with AutoTrain" (which is highlighted), "conversational", "Inference Endpoints", and "License: other". The main content area has a tab bar with "Model card" (selected), "Files and versions", "Training metrics", "Community", and "Settings". A button labeled "Edit model card" is visible. The "Model Trained Using AutoTrain" section contains the text: "This model was trained using AutoTrain. For more information, please visit [AutoTrain](#)". The "Usage" section shows a code snippet:

```
from transformers import AutoModelForCausalLM, AutoTokenizer  
  
model_path = "PATH_TO_THIS_REPO"
```

On the right side, there's a sidebar with "Downloads last month" (0) and a "Text Generation" section stating "Unable to determine this model's library. Check the".

# Under the hood

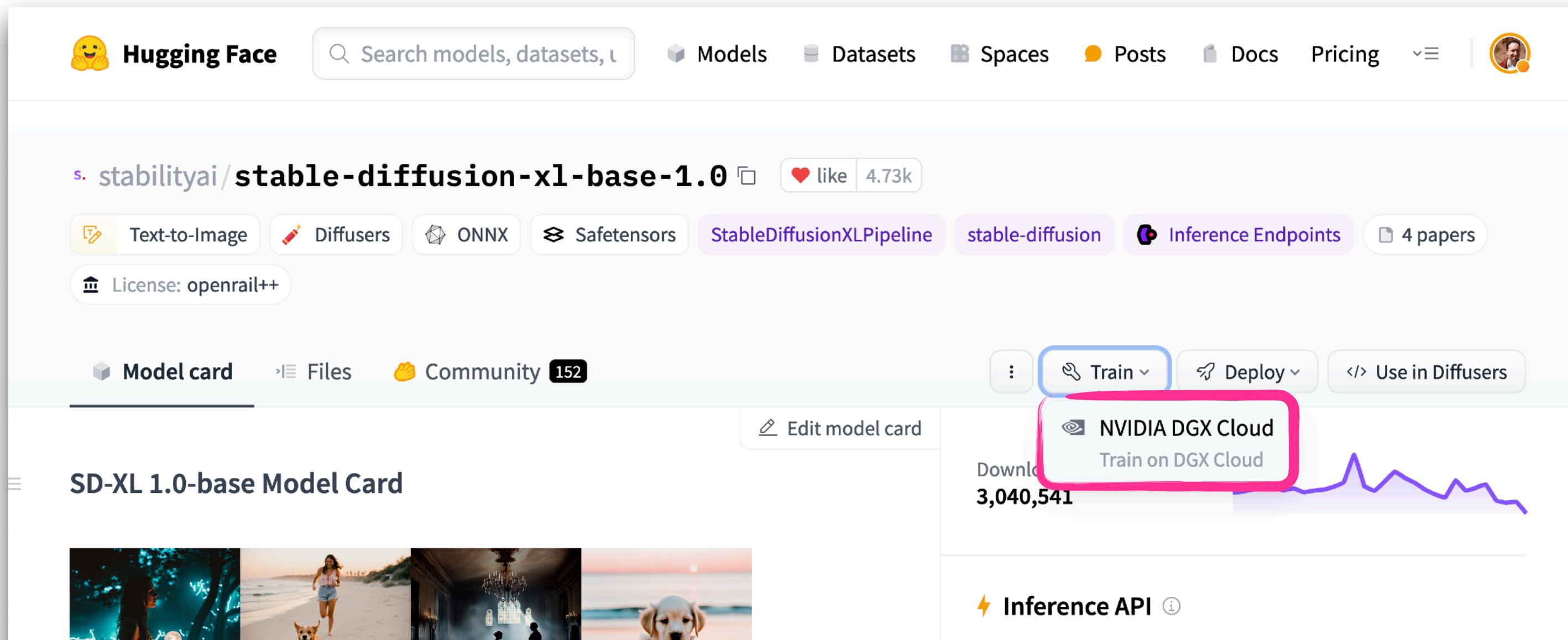


Within your Enterprise Hub org,  
a Space is created,  
hosting AutoTrain.  
You create training jobs,  
they execute on DGX Cloud,  
private models are saved,  
you pay for used compute.

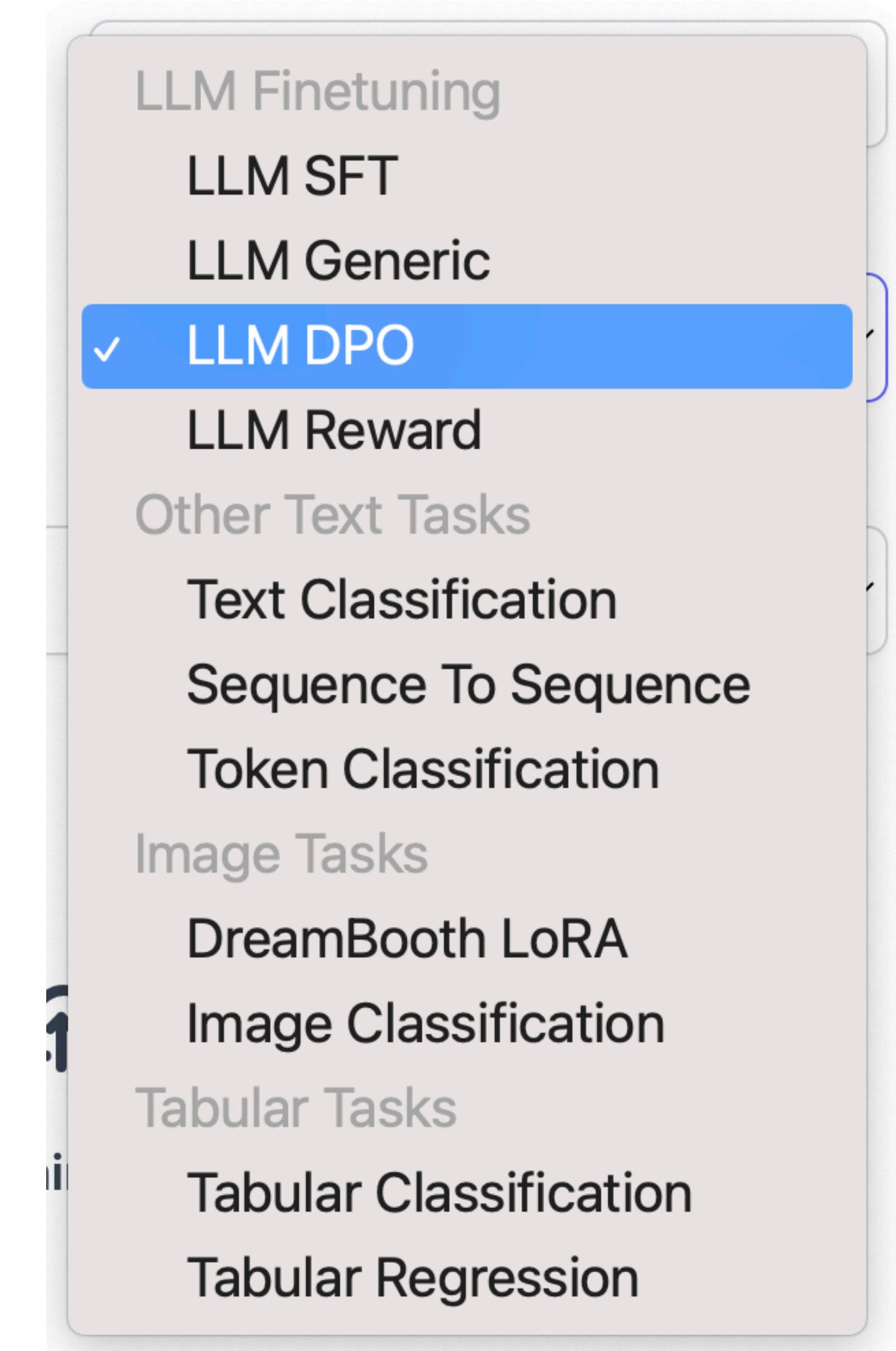
# Train your own models

 **LLMs: SFT, DPO alignment, Reward**  
**Llama 2, Mistral, Mixtral, Gemma...**

 **Image generation: Dreambooth Lora**  
**Stable Diffusion, Stable Diffusion XL**



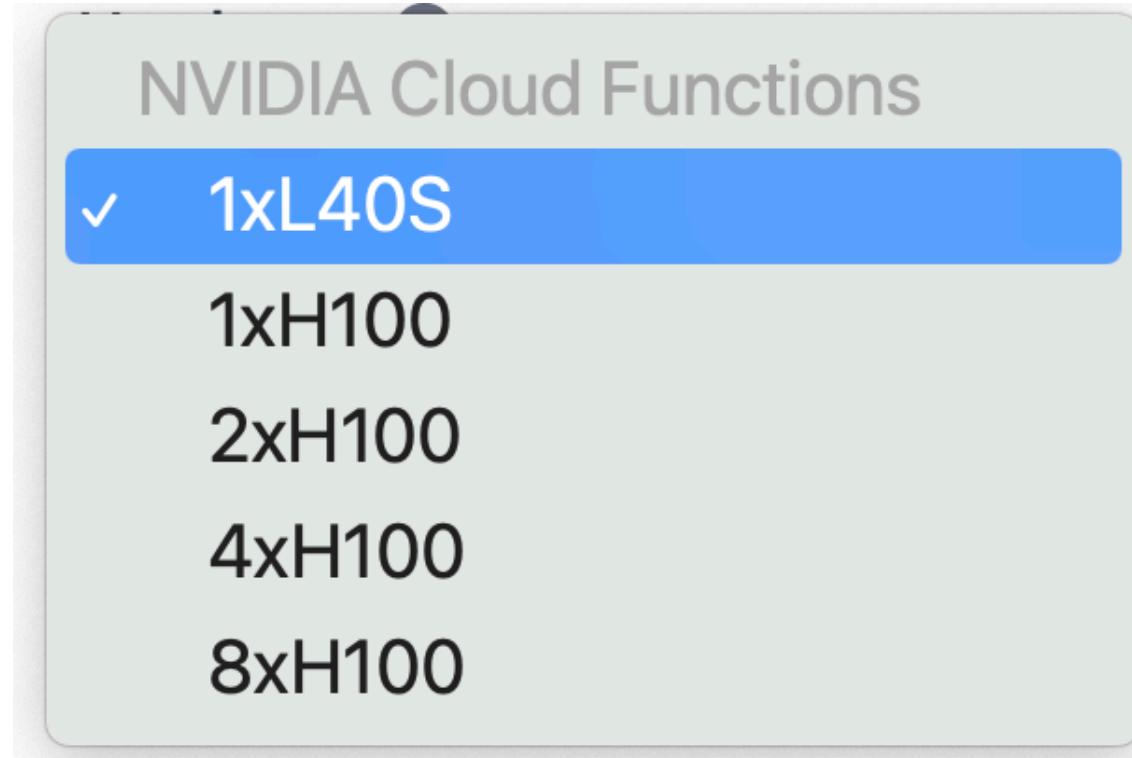
The screenshot shows the Hugging Face Model Hub interface. At the top, there's a navigation bar with links for Hugging Face, Search models, datasets, etc., and a user profile. Below the search bar, a model card for "stabilityai/stable-diffusion-xl-base-1.0" is displayed. The card includes tabs for Model card, Files, and Community (with 152 contributions). It features a "Train" button, which is highlighted with a red box and has a tooltip "NVIDIA DGX Cloud Train on DGX Cloud". Other buttons include Deploy, Use in Diffusers, and Edit model card. The card also shows download statistics (3,040,541) and a line graph. At the bottom, there are preview images of generated images.



The sidebar lists various training and modeling tasks:

- LLM Finetuning
- LLM SFT
- LLM Generic
- LLM DPO** (selected)
- LLM Reward
- Other Text Tasks
- Text Classification
- Sequence To Sequence
- Token Classification
- Image Tasks
- DreamBooth LoRA
- Image Classification
- Tabular Tasks
- Tabular Classification
- Tabular Regression

# Pay for used compute



NVIDIA GPU	GPU Memory	Price / hour
H100	80 GB	\$8.25
L40S	48 GB	\$2.75

*For example, fine-tuning Mistral 7B on 1500 samples on a single NVIDIA L40S takes ~10 minutes and costs ~\$0.45*

# Introducing...

# optimum-nvidia



# One line code change...

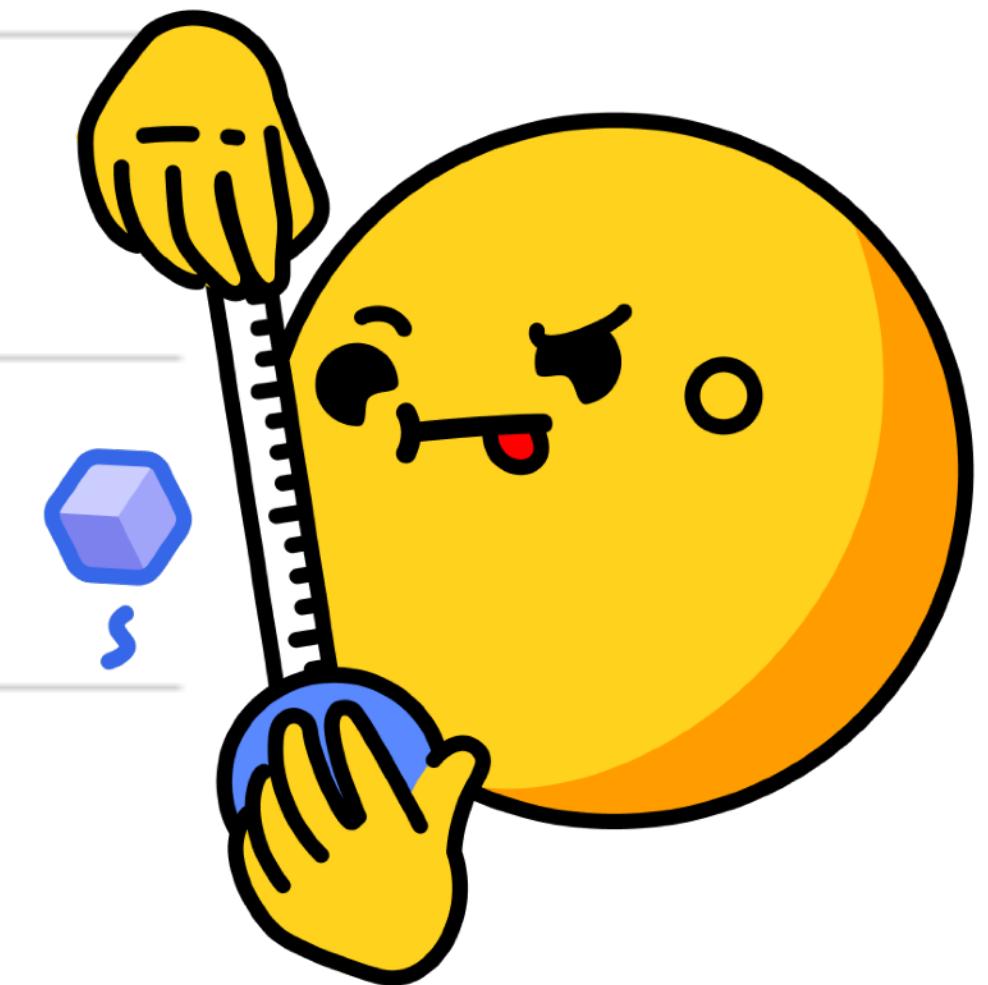
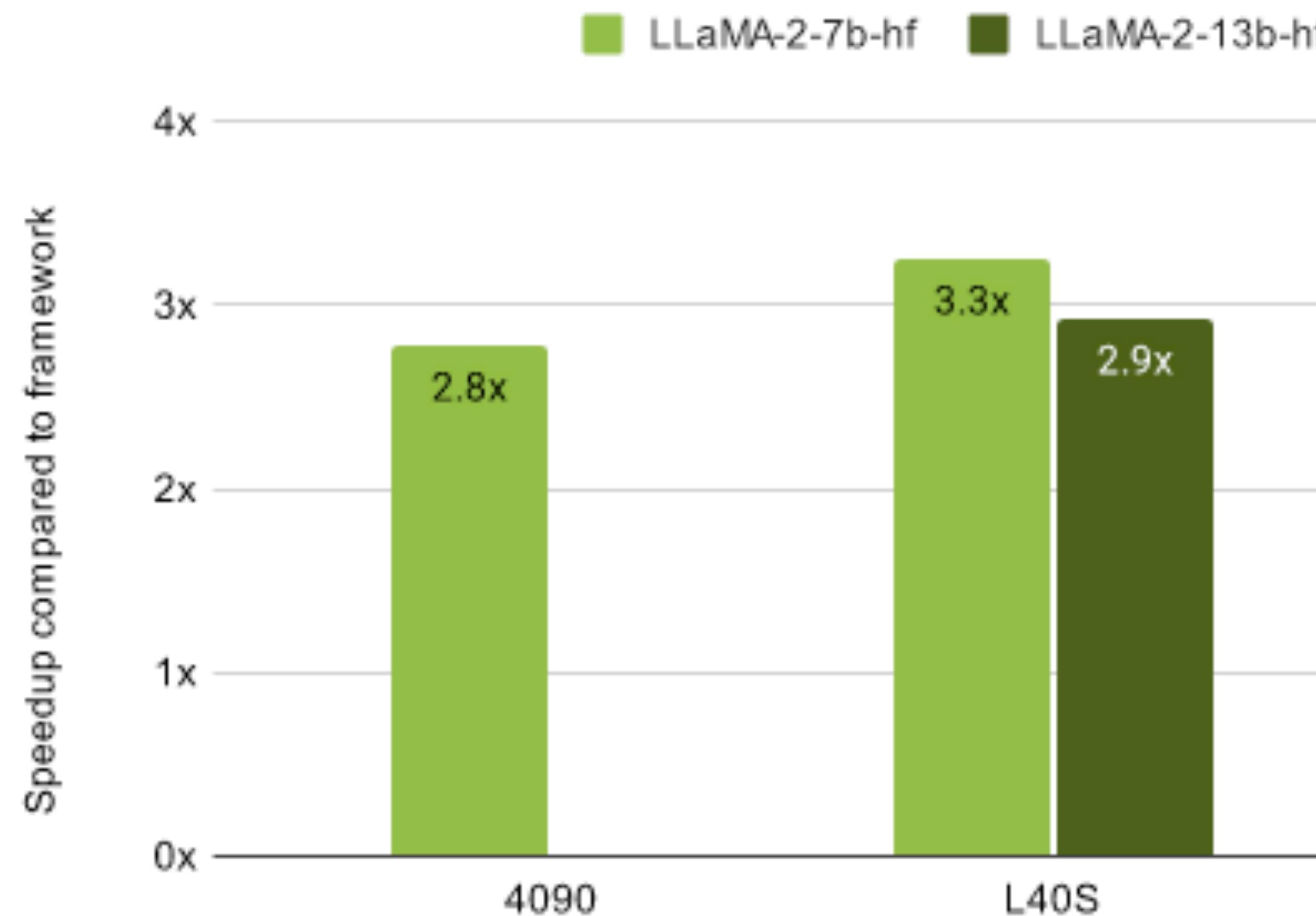
```
- from transformers.pipelines import pipeline  
+ from optimum.nvidia.pipelines import pipeline  
  
# everything else is the same as in transformers!  
pipe = pipeline('text-generation', 'meta-llama/Llama-2-7b-chat-hf', use_fp8=True)  
pipe("Describe a real-world application of AI in sustainable energy.")
```

... to leverage FP8 with TensorRT-LLM

# Get 3x First Token Latency

LLaMA 2 First Token Latency

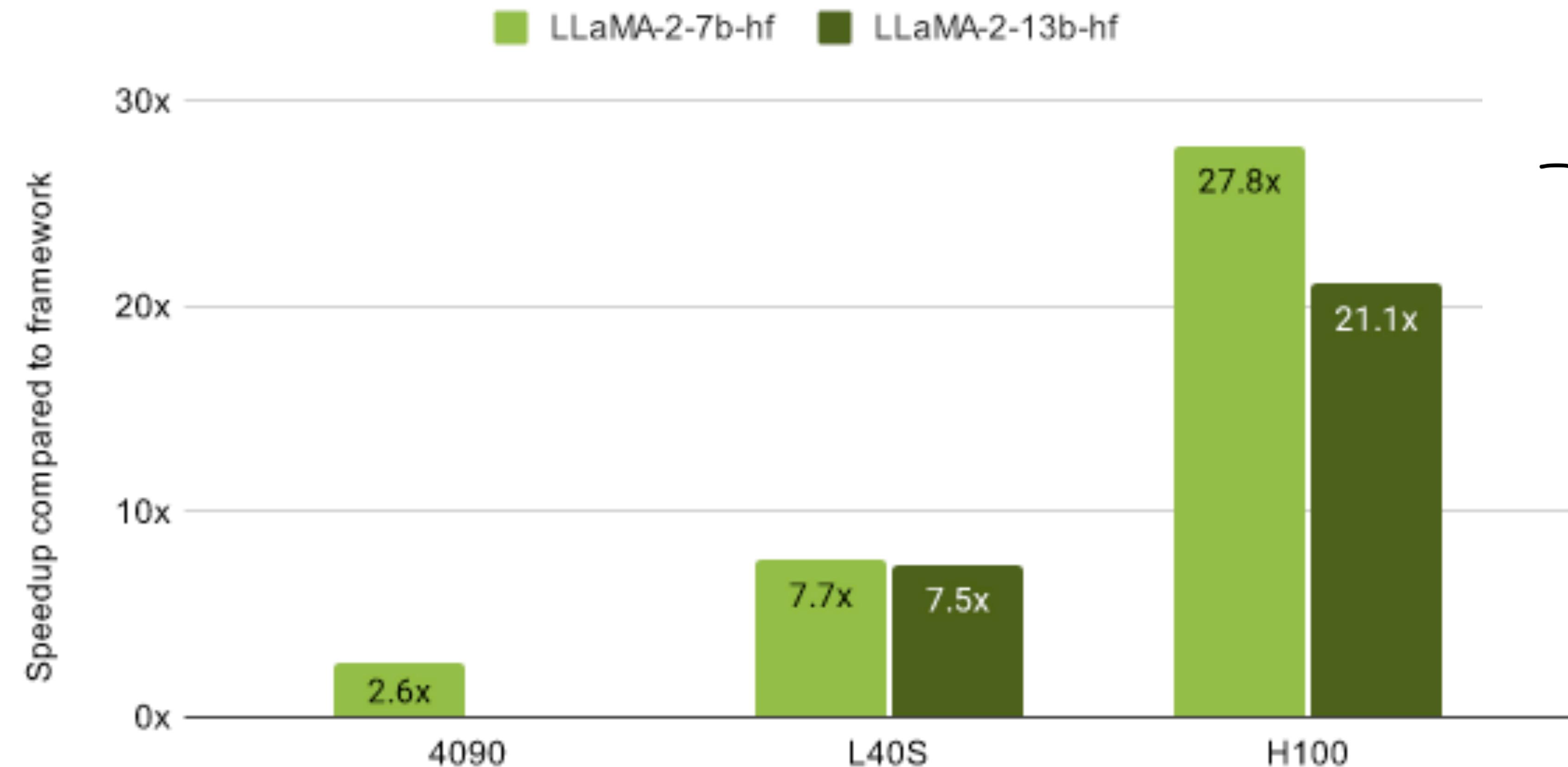
Batch Size = 1



# And up to 28x throughput!

LLaMA 2 Throughput

Batch Size = 4



# We're just getting started!

**Train on DGX Cloud**

**optimum-nvidia**

**NIM Microservices**

**NeMo Data Curator**

**>100 models on [hf.co/nvidia](https://hf.co/nvidia)**

# Building Accelerated AI with Hugging Face + NVIDIA

GTC March 2024

Jeff Boudier, Product @ 





Thank you!