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

6 things about

you probably didn't know

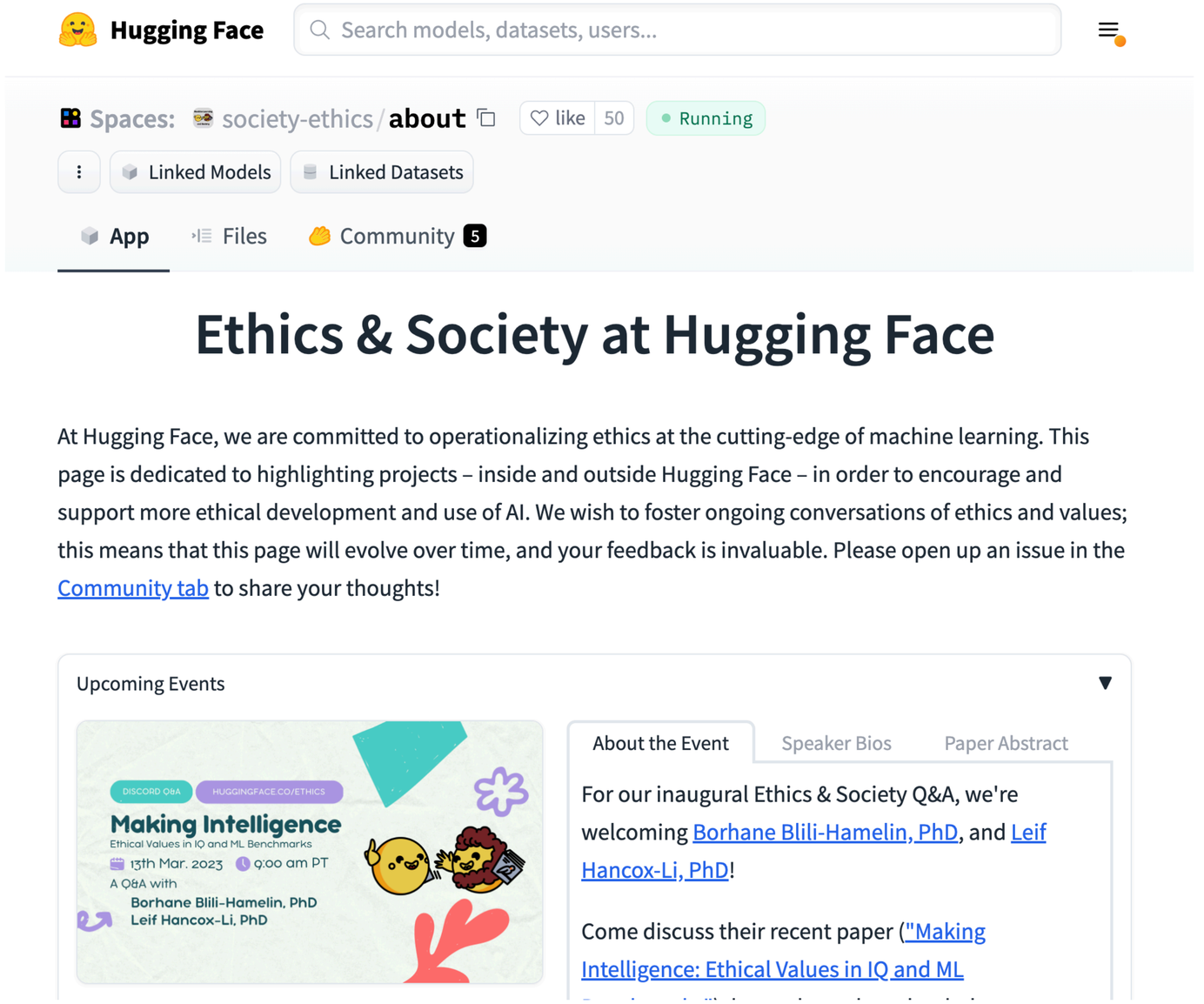
Building Accelerated AI with Hugging Face + NVIDIA - GTC March 2024

1️⃣ Hugging Face

Democratize Good Machine Learning

open source, community, ethics-first

2 Hugging Face



hf.co/ethics
hf.co/blog/ethics-diffusers

3 Hugging Face

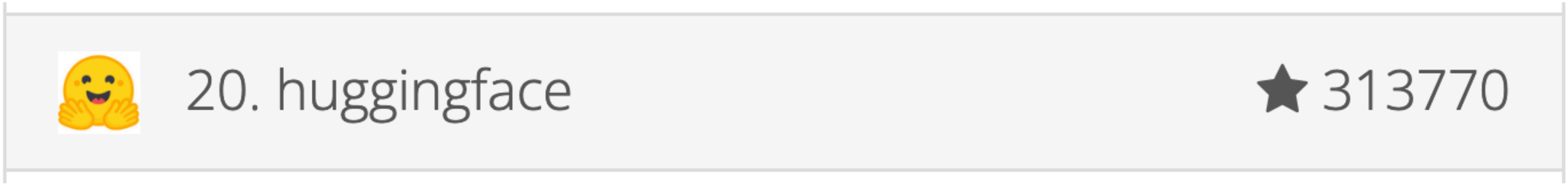
transformers

datasets

peft

diffusers

tokenizers



text-generation-inference

optimum

timm

trl

chat-ui

evaluate

nanotron

accelerate

setfit

safetensors

candle

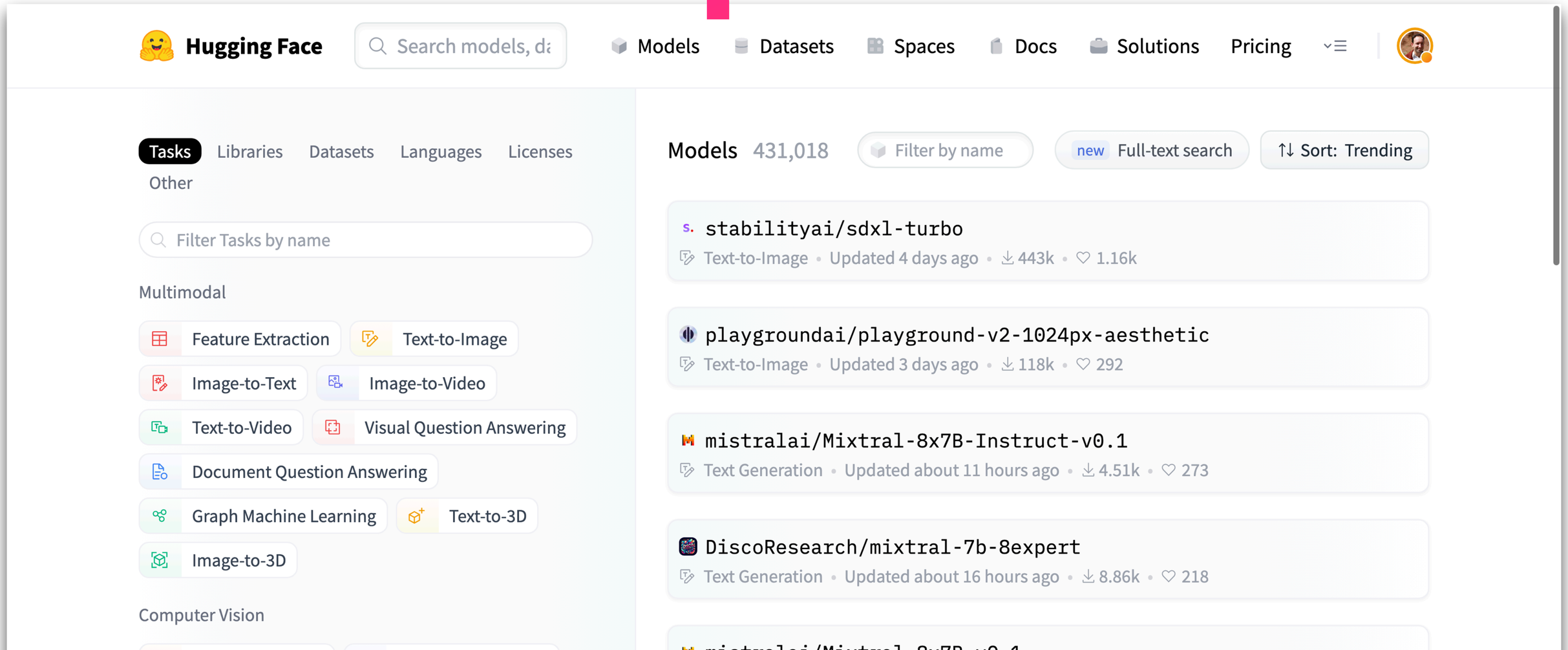
datatrove

lighteval

Source: gitstar-ranking.com

4 Hugging Face

>500k free public models



5 Hugging Face

/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

- **Depth Estimation** — 47 models
- **Image Classification** — 2,260 models
- **Image Segmentation** — 169 models
- **Image-to-Image** — 48 models
- **Object Detection** — 284 models
- **Video Classification** — 73 models
- **Unconditional Image Generation** — 533 models
- **Zero-Shot Image Classification** — 63 models

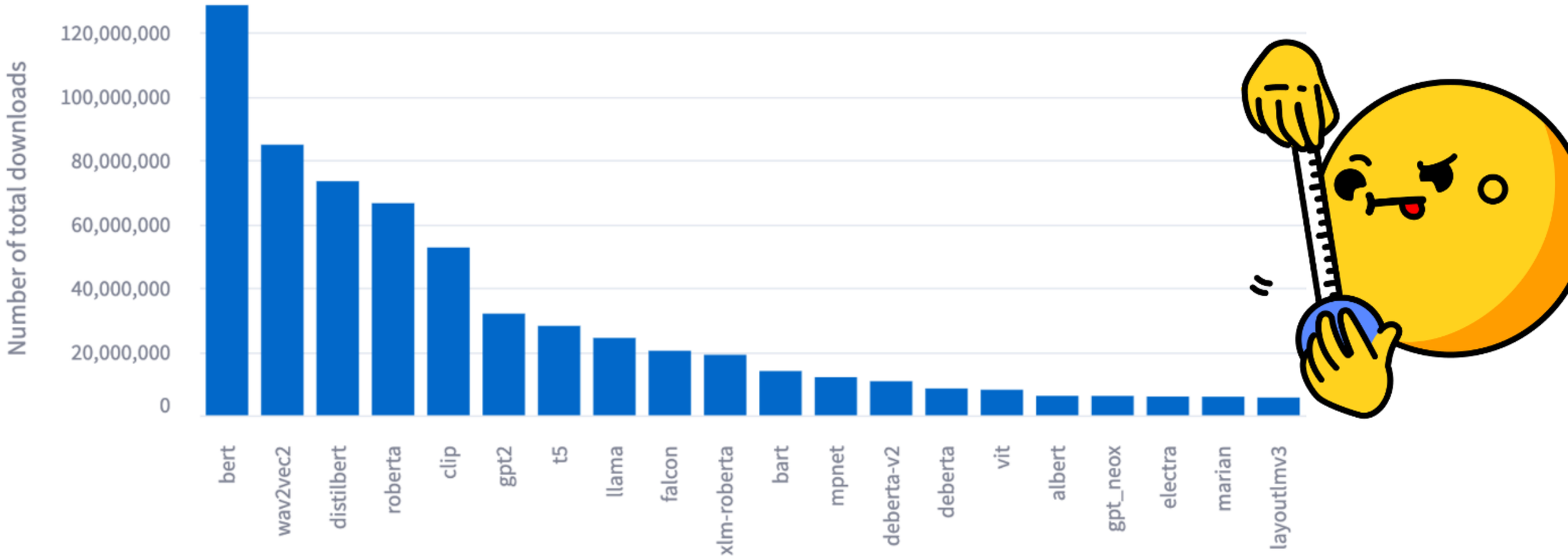
Natural Language Processing

- **Conversational** — ,924 models
- **Fill-Mask** — 5,926 models
- **Question Answering** — 3,862 models
- **Sentence Similarity** — 1,558 models
- **Summarization** — 873 models
- **Table Question Answering** — 56 models
- **Text Classification** — 17,572 models
- **Text Generation** — 8,631 models

Source: hf.co/tasks

6 Hugging Face

>10M model downloads /day

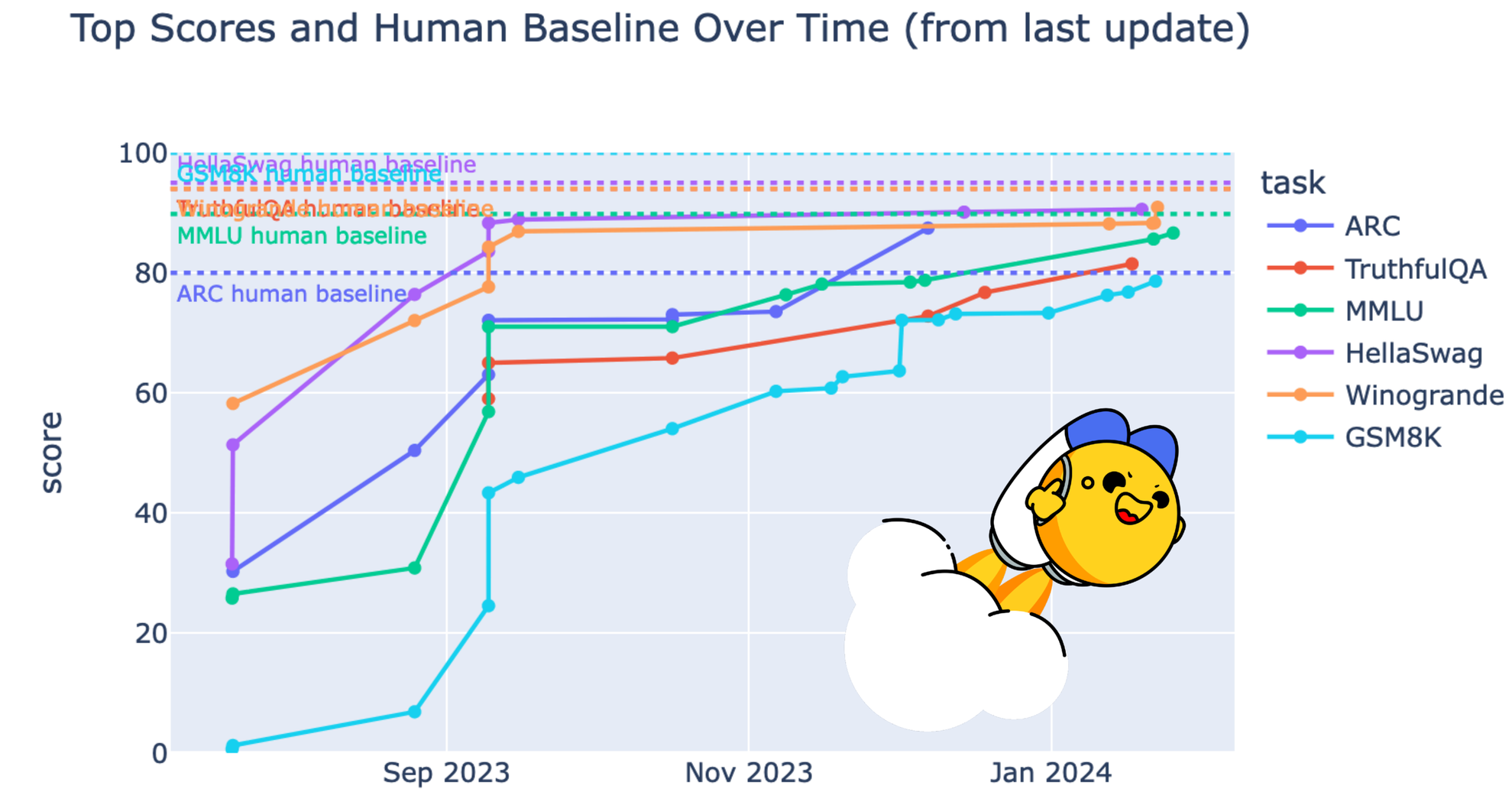


Transformers model downloads, last 30 days
Source: hf.co/spaces/huggingface/transformers-stats

Why you should build AI with open source

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

1️⃣ Future-proof: open models 🚀



Source: hf.co/spaces/HuggingFaceH4/open_llm_leaderboard

2️⃣ 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

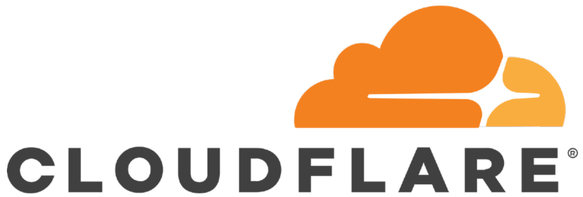
Faster Stable Diffusion with Core ML on iPhone, iPad, and Mac

By pcuenq ▪ June 15, 2023

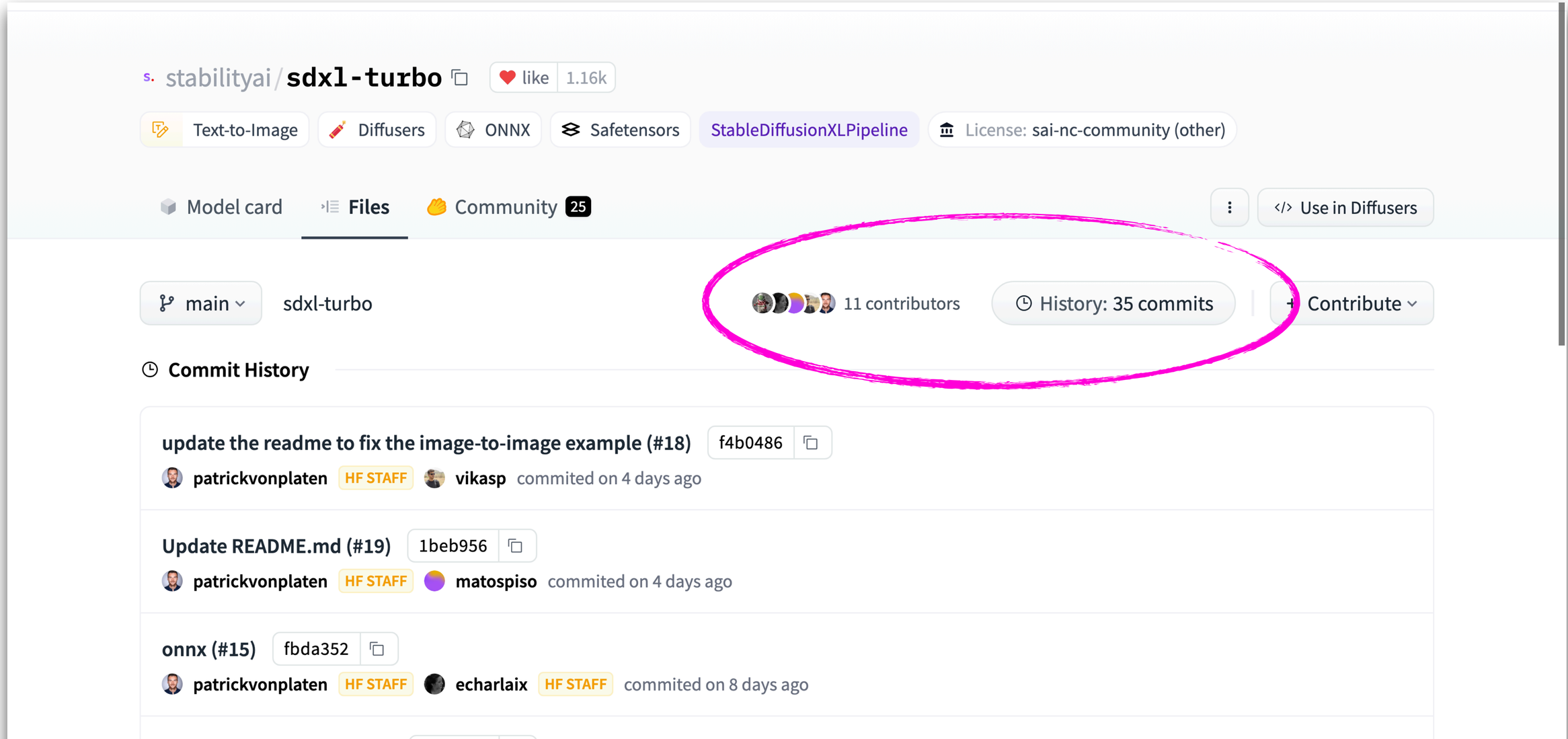
SDXL in 4 steps with Latent Consistency LoRAs

By pcuenq ▪ November 9, 2023

3️⃣ Security: Your model, your hosting 🔒

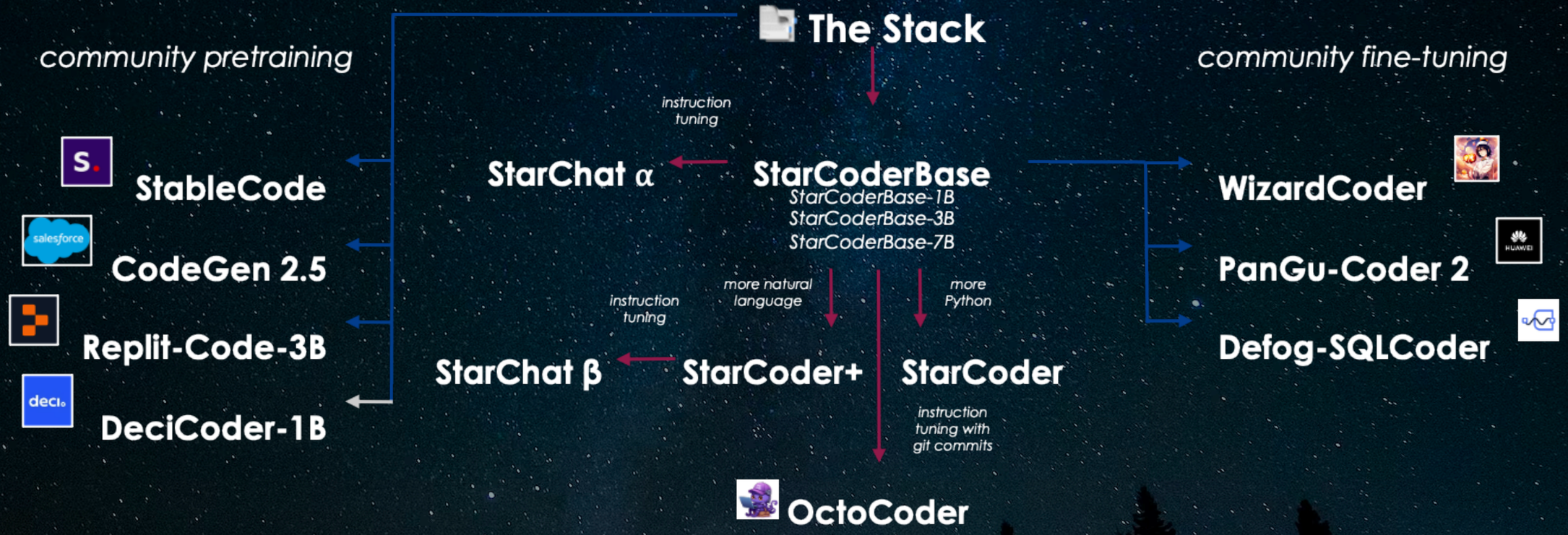


4️⃣ Models you (version) control 🎛️



5 Models you can trust (and verify)

BigCode Ecosystem

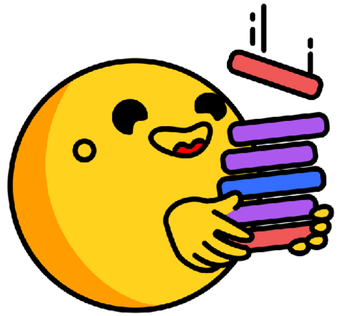


Why build AI with open source?

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

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

LLMs: Compare them all!

Open LLM Leaderboard

The Open LLM Leaderboard aims to track, rank and evaluate open LLMs and chatbots

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 ressour

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

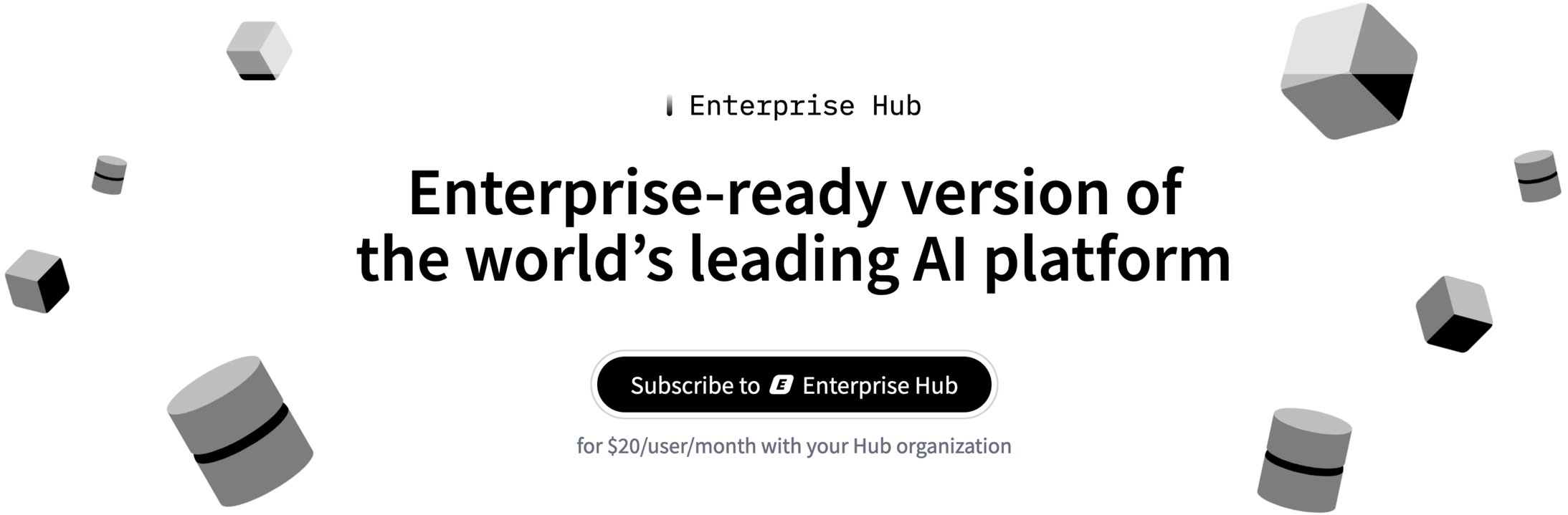
Filter model types

all | 🟢 pretrained | 🔶 fine-tuned | ⭕ instructi

🟦 RL-tuned

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

100,000 organizations



Enterprise Hub

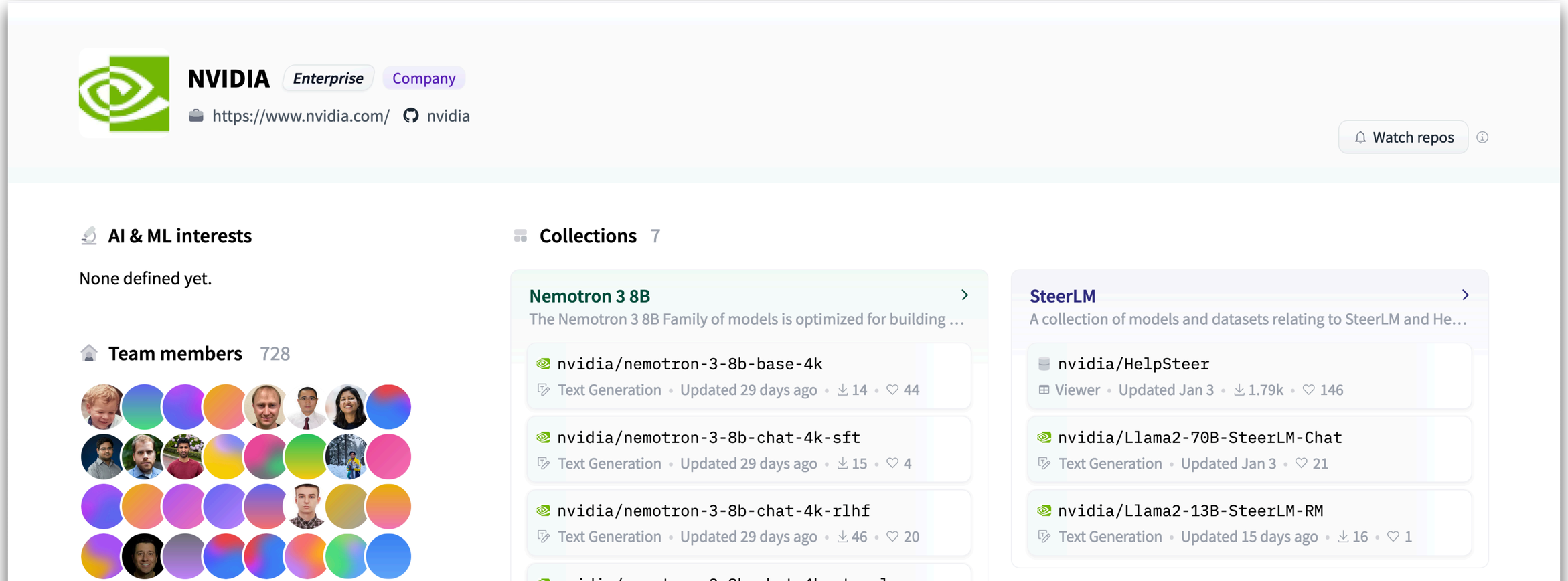
Enterprise-ready version of the world's leading AI platform

Subscribe to Enterprise Hub

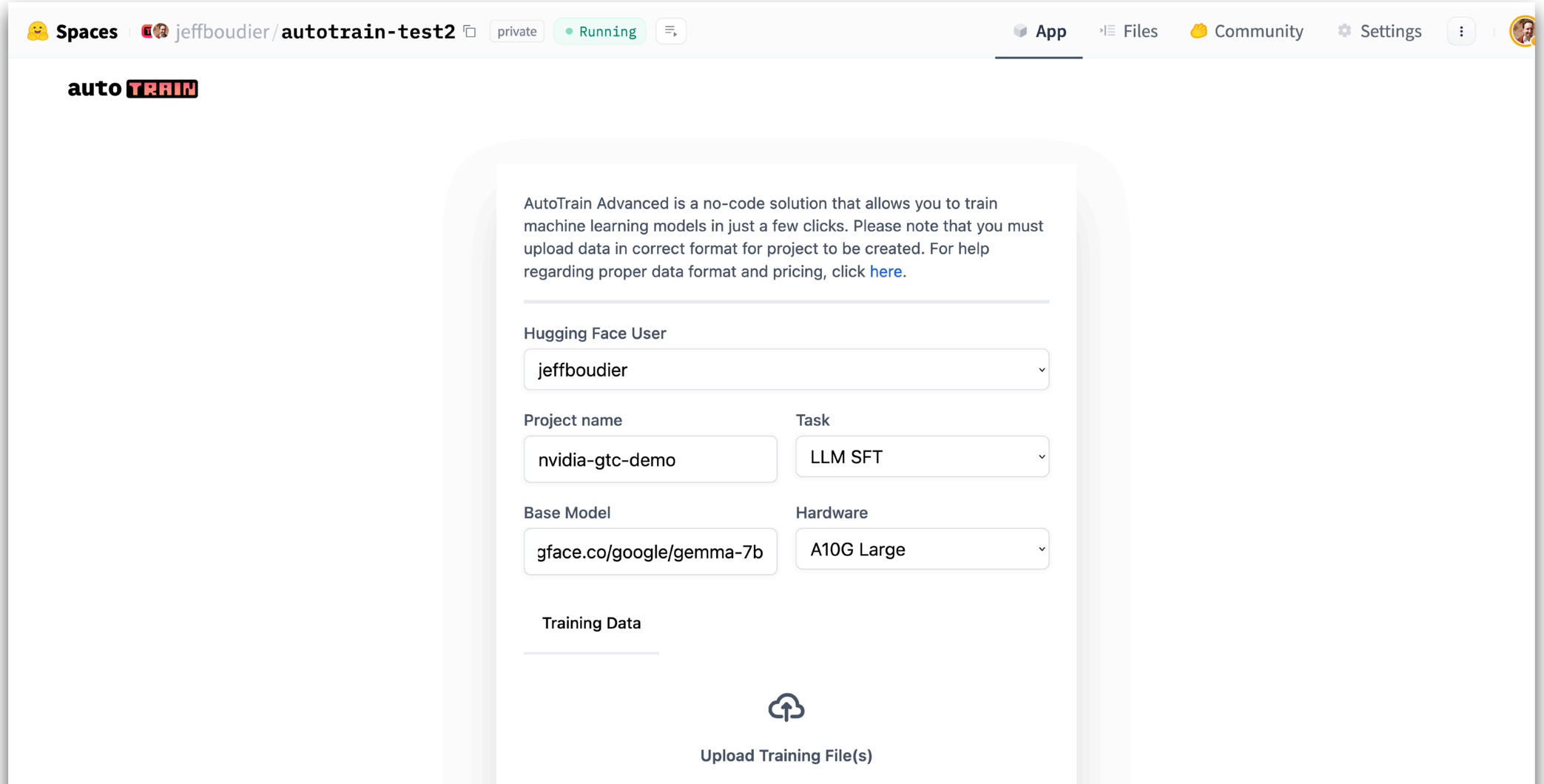
for $20/user/month with your Hub organization

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



Train it easy with AutoTrain



text-generation-inference

- Tensor Parallelism
- Token Streaming
- Metrics and monitoring
- Quantization
- Optimizations
- 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.

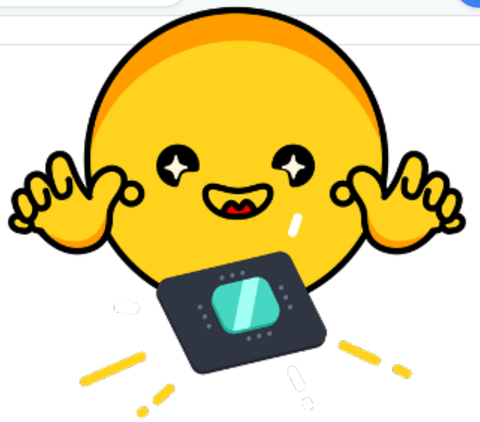
GPU 2x Nvidia A100 $ 13 / h Go

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

Spaces

Discover amazing ML apps made by the community!

Create new Space or learn more about Spaces.

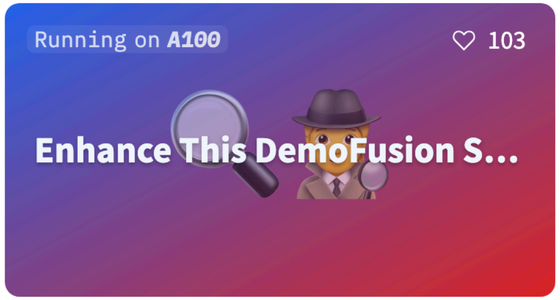
Search Spaces

new Full-text search

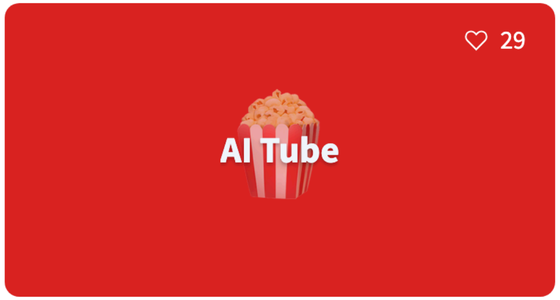
↑↓ Sort: Trending

☆ Spaces of the week 🔥

playgroundai 6 days ago

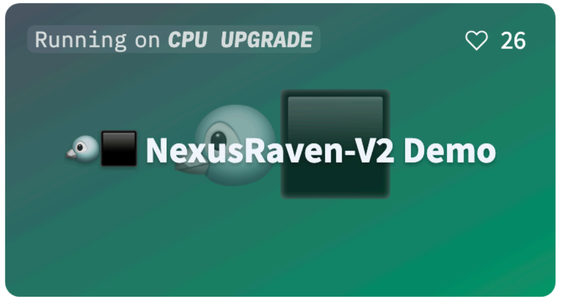


radames about 3 hours ago

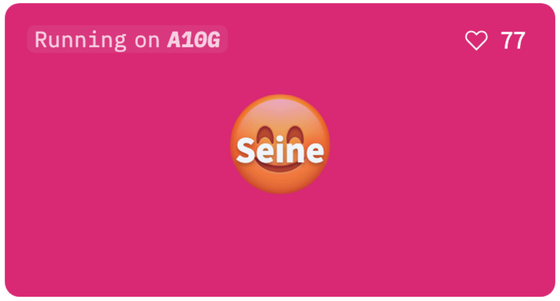


jbilcke-hf about 16 hours ago

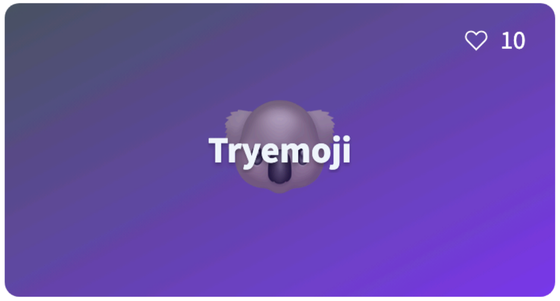
toshas 4 days ago



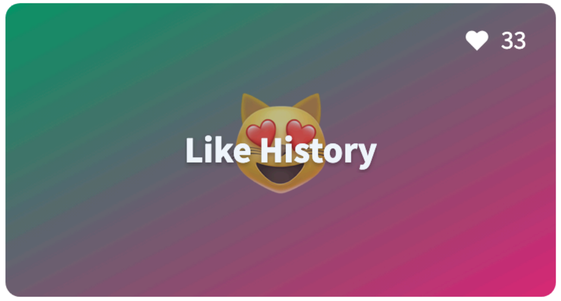
Nexusflow 6 days ago



Vchitect 5 days ago

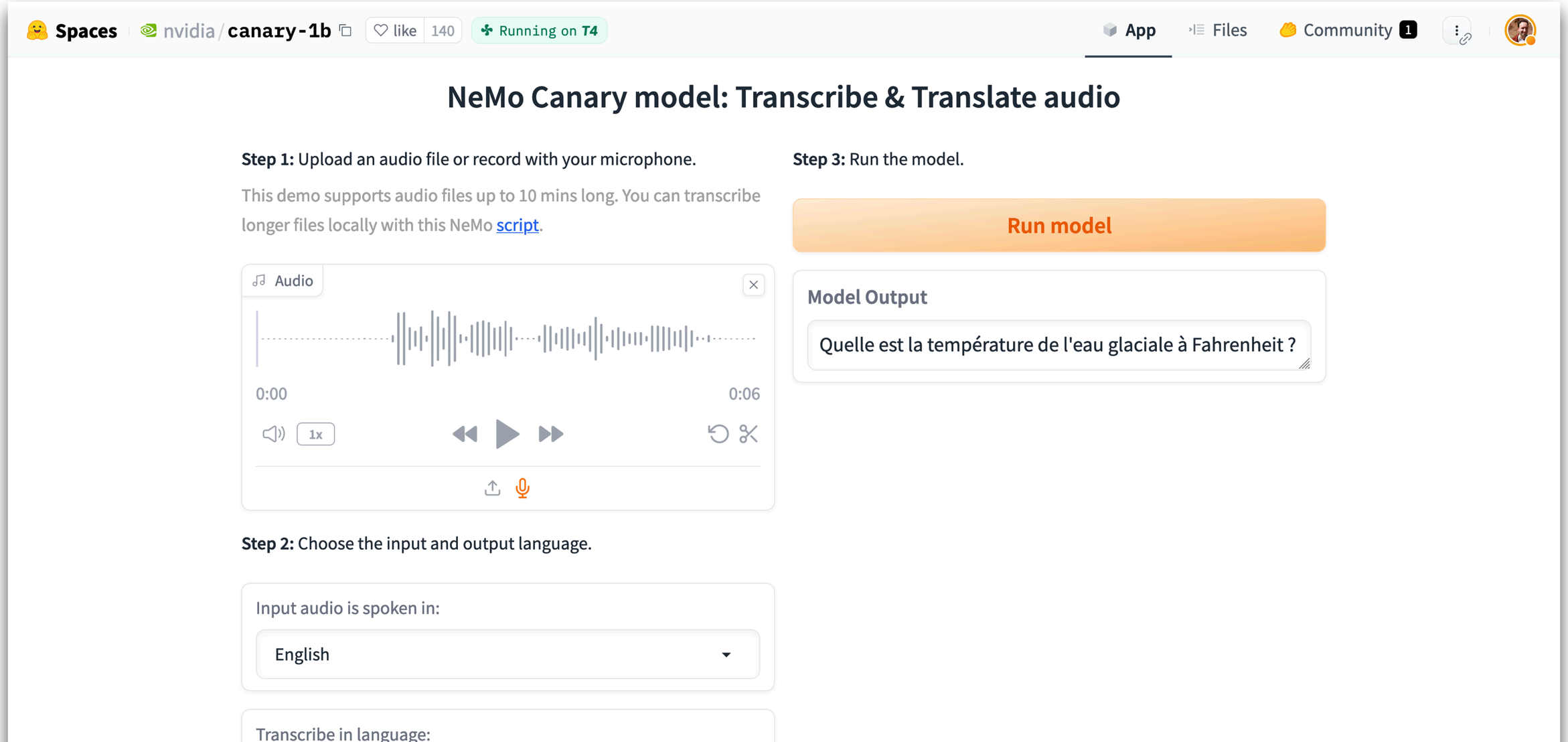


leptonai 12 days ago



timqian 5 days ago

500,000 Spaces created



Accelerate Spaces with NVIDIA

Space Hardware

Display price: per hour per month

Choose a hardware for your Space.

You'll be billed on a per minute basis.
View usage in your billing settings.

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

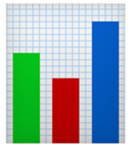
AI Accelerator
HPU · IPU · ...

Sleep time settings

Sleep after 24 hours of inactivity

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

How to build AI with Hugging Face

 Assess models with Leaderboards

 Collaborate with Enterprise Hub

 Train with AutoTrain

 Deploy with Inference Endpoints

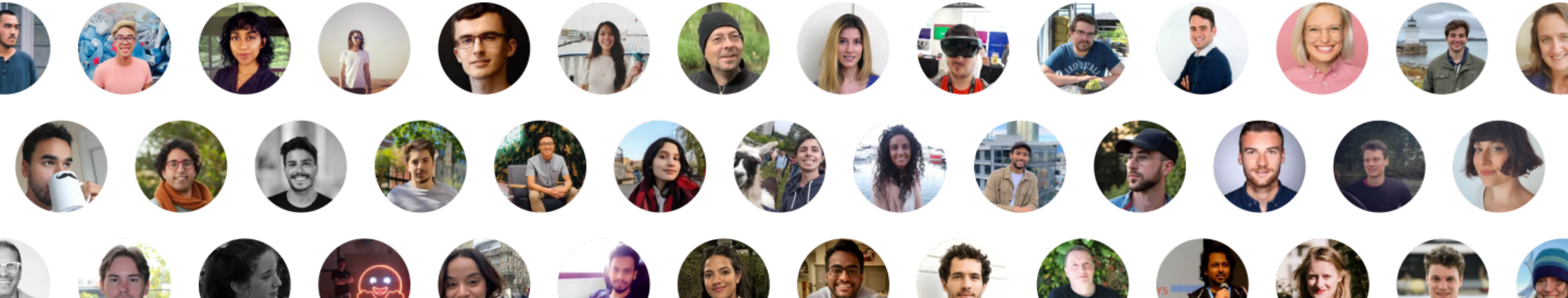
 Demo with Spaces

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

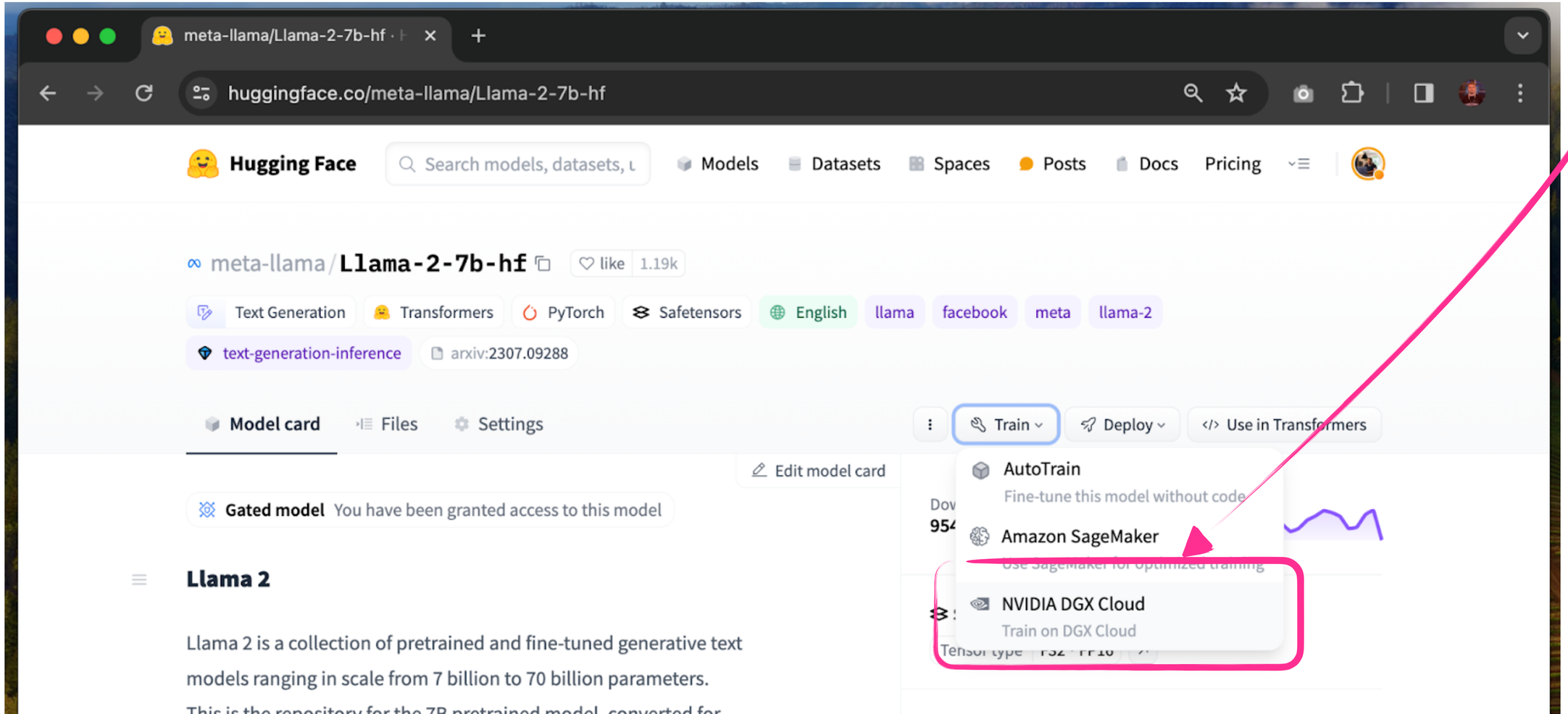
Introducing...

Train on DGX Cloud

Finetune LLMs on the Hub with H100 and L40S without code

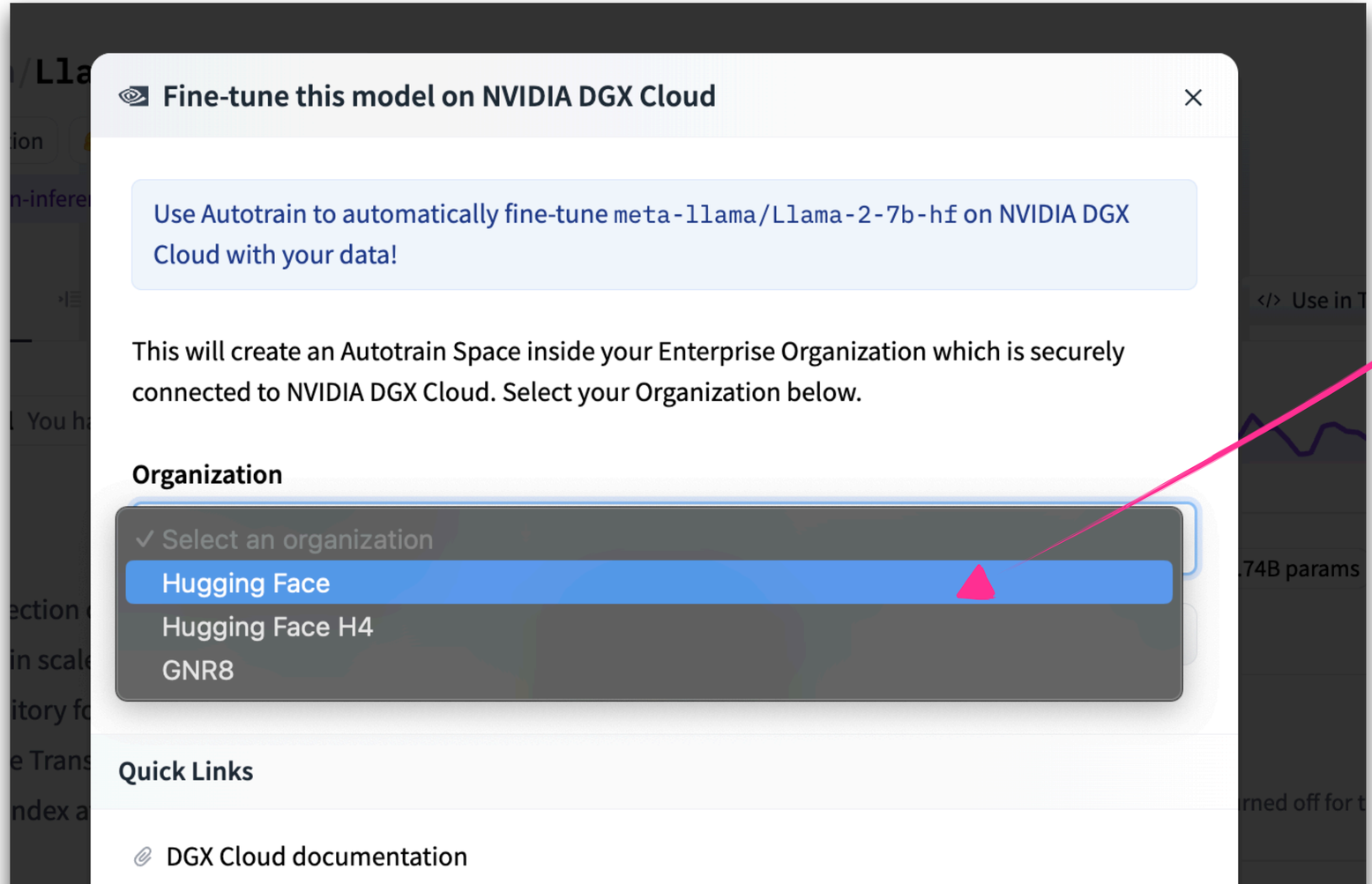
How it works

1. Pick a model, click Train on DGX Cloud



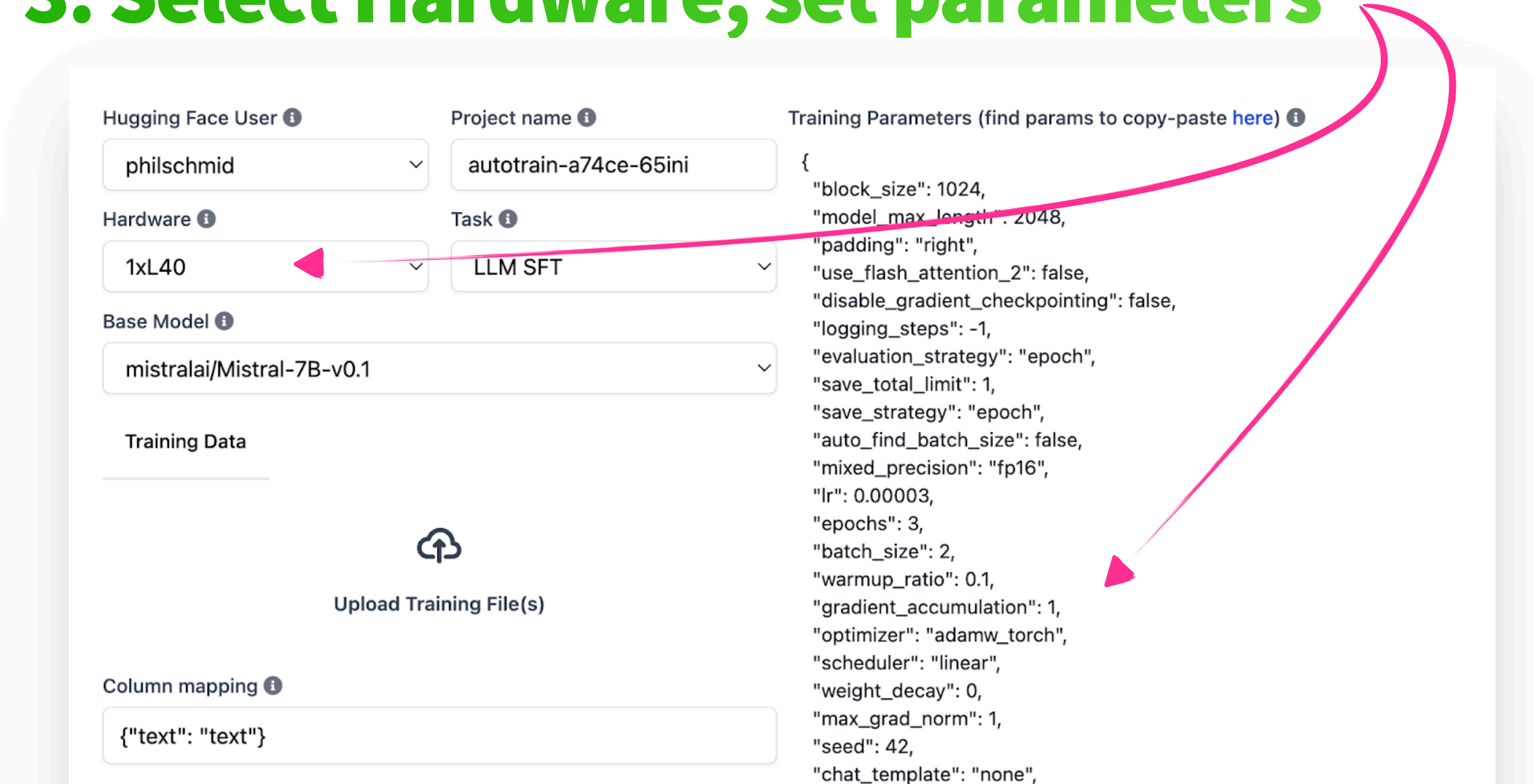
How it works

2. Select your Enterprise Hub organization

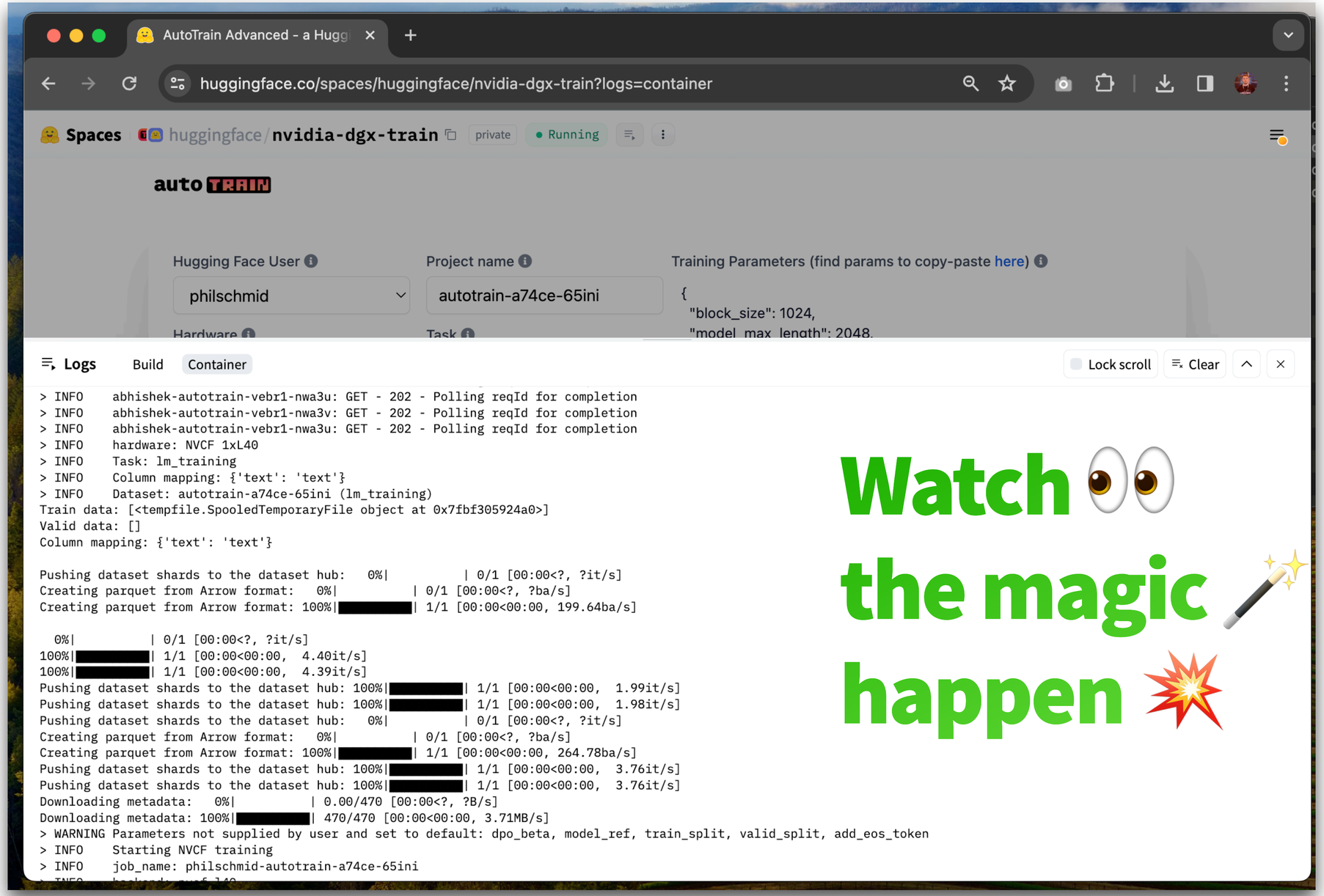


How it works

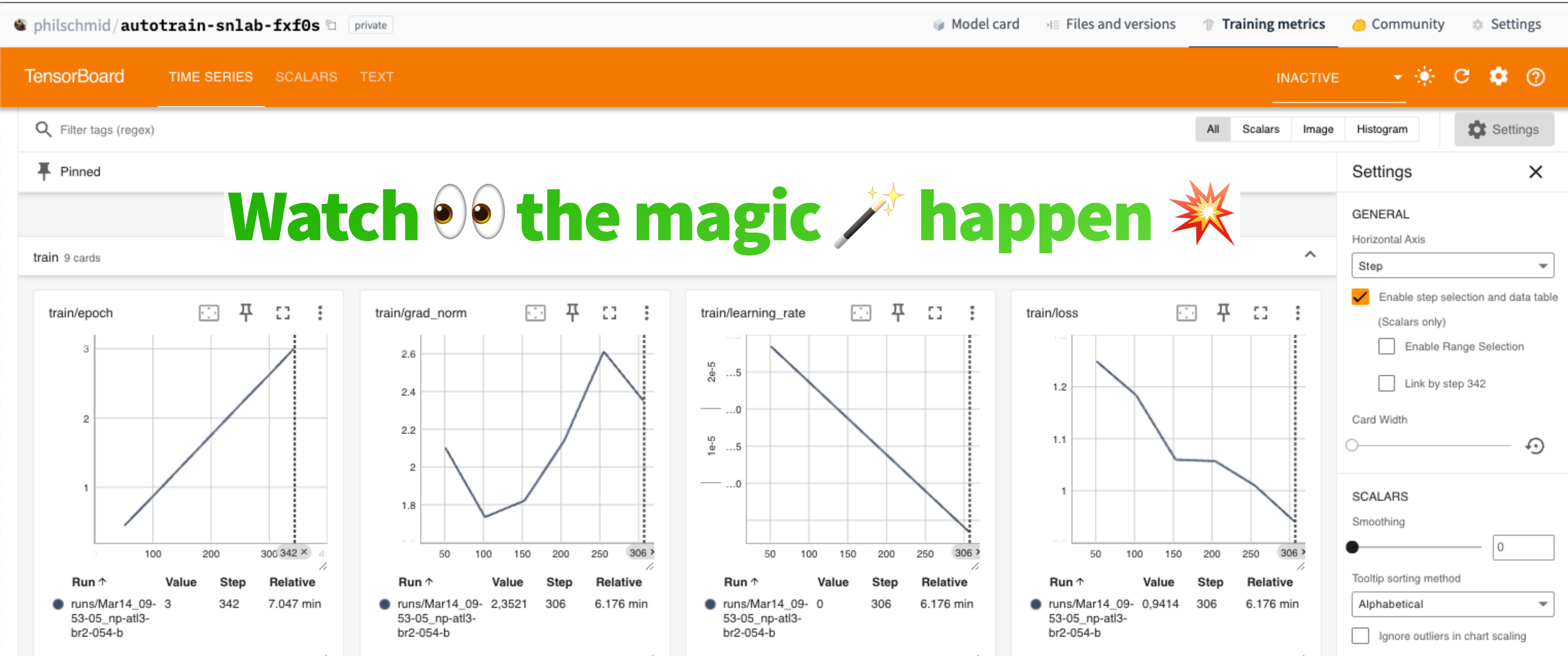
3. Select Hardware, set parameters



How it works

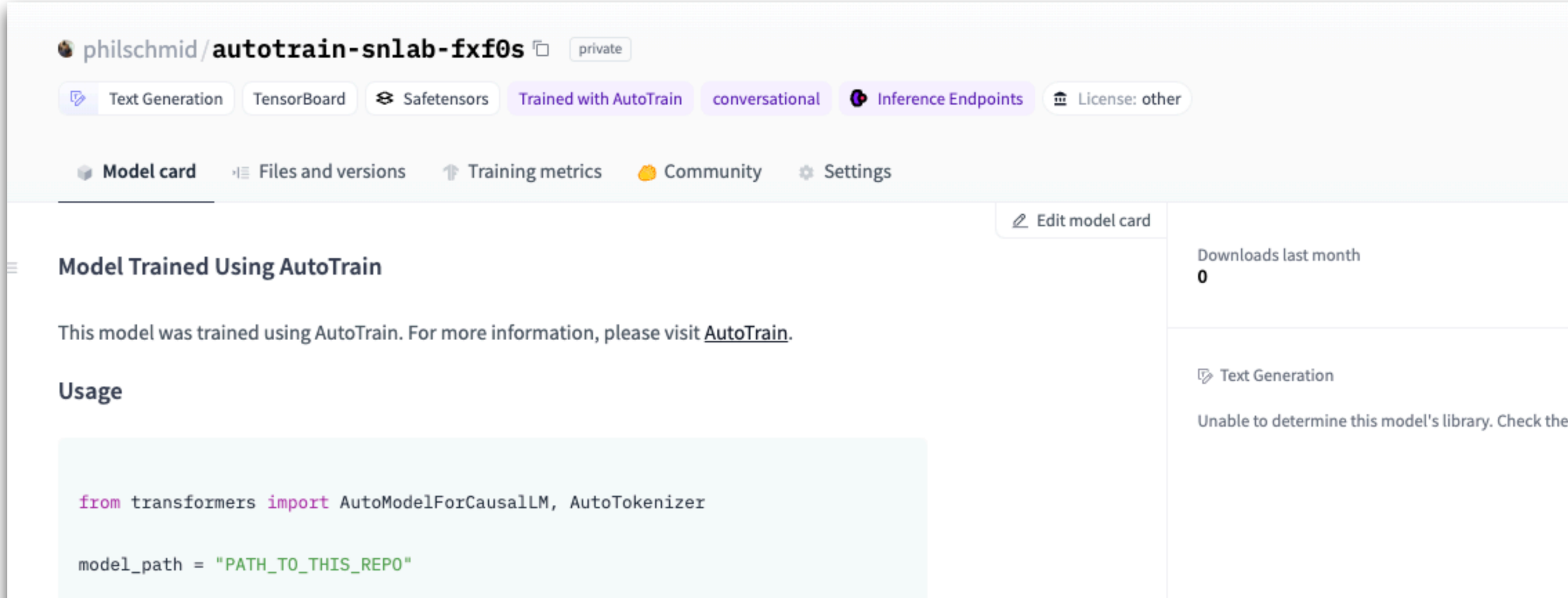

AutoTrain Advanced - a Hugg
huggingface.co/spaces/huggingface/nvidia-dgx-train?logs=container
Spaces
huggingface / nvidia-dgx-train
private
Running
auto TRAIN
Hugging Face User
philschmid
Project name
autotrain-a74ce-65ini
Training Parameters (find params to copy-paste here)
{
"block_size": 1024,
"model_max_length": 2048,
Hardware
Task
Logs
Build
Container
Lock scroll
Clear
> 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
Watch 👀
the magic ✨
happen 💥

How it works

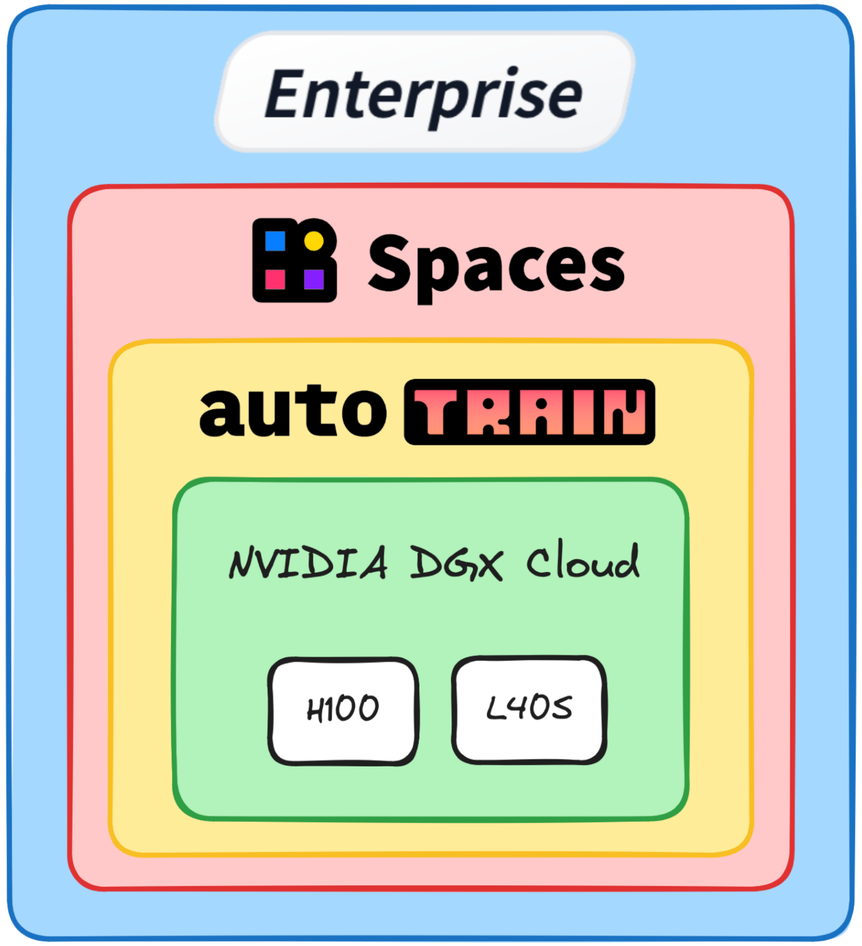

philschmid / autotrain-snlab-fxf0s
private
Model card
Files and versions
Training metrics
Community
Settings
TensorBoard
TIME SERIES
SCALARS
TEXT
INACTIVE
Filter tags (regex)
All
Scalars
Image
Histogram
Settings
Pinned
Watch 👀 the magic 🪄 happen 💥
train 9 cards
train/epoch
train/grad_norm
train/learning_rate
train/loss
Run ↑ Value Step Relative
runs/Mar14_09-53-05_np-atl3-br2-054-b 3 342 7.047 min
runs/Mar14_09-53-05_np-atl3-br2-054-b 2,3521 306 6.176 min
runs/Mar14_09-53-05_np-atl3-br2-054-b 0 306 6.176 min
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
0
Tooltip sorting method
Alphabetical
Ignore outliers in chart scaling

How it works

4. Check out your new private model!



Under the hood

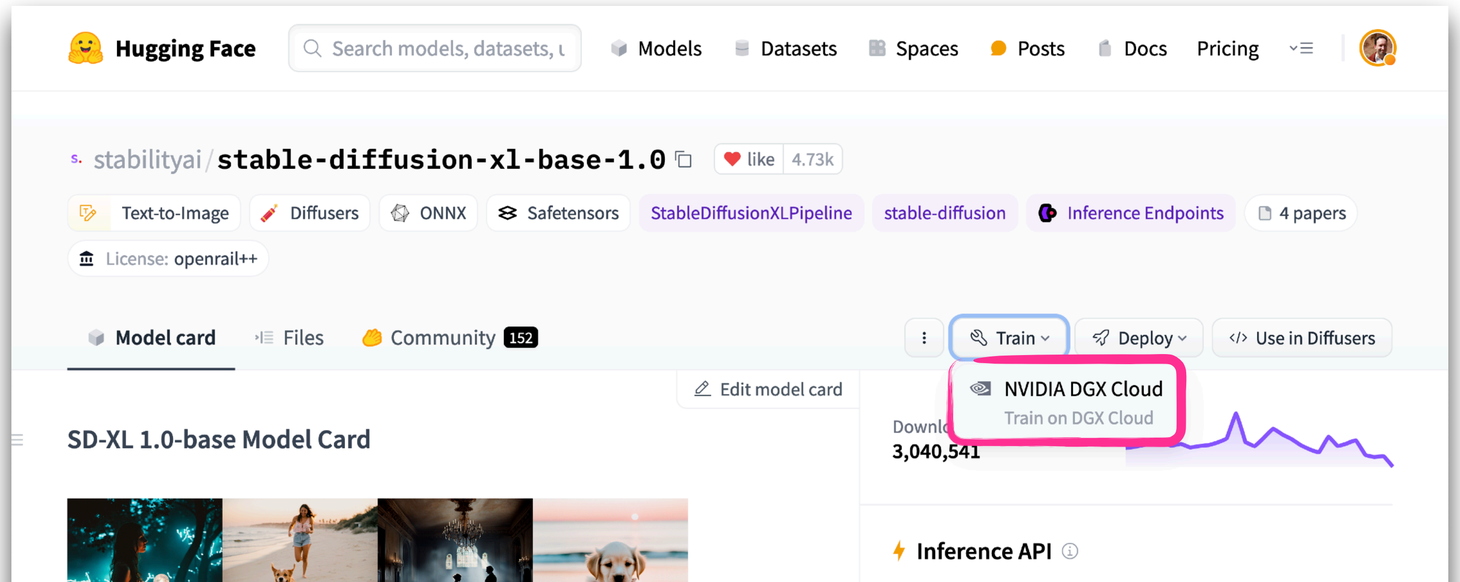


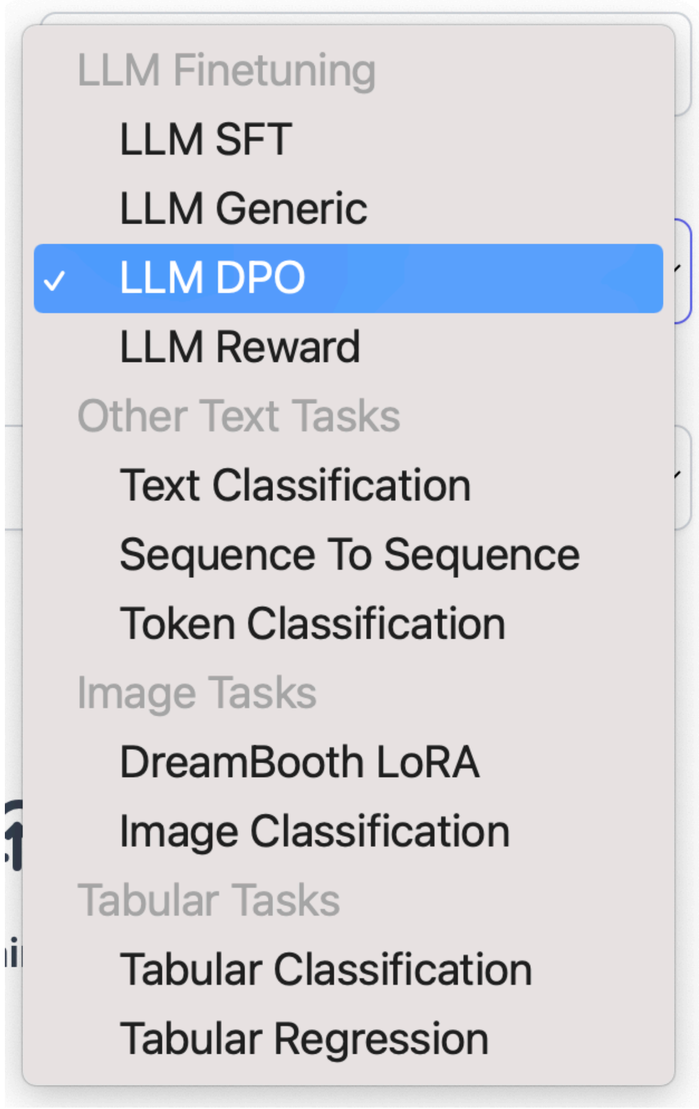
Within your Enterprise Hub org,
a Space is created,
hosting AutoTrain.
You create training jobs,
they execute on DGX Cloud,
private models are saved,
you pay for used compute.

Train your own models

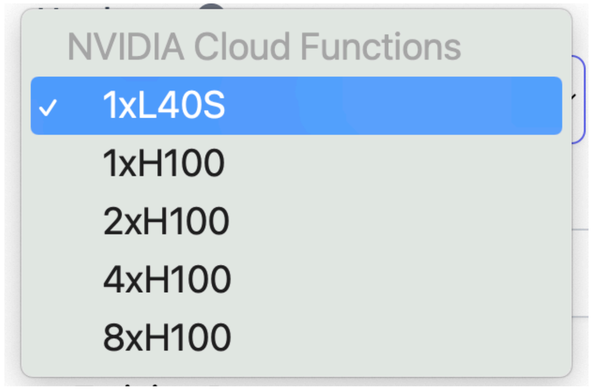
LLMs: SFT, DPO alignment, Reward
Llama 2, Mistral, Mixtral, Gemma…

Image generation: Dreambooth Lora
Stable Diffusion, Stable Diffusion XL





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...

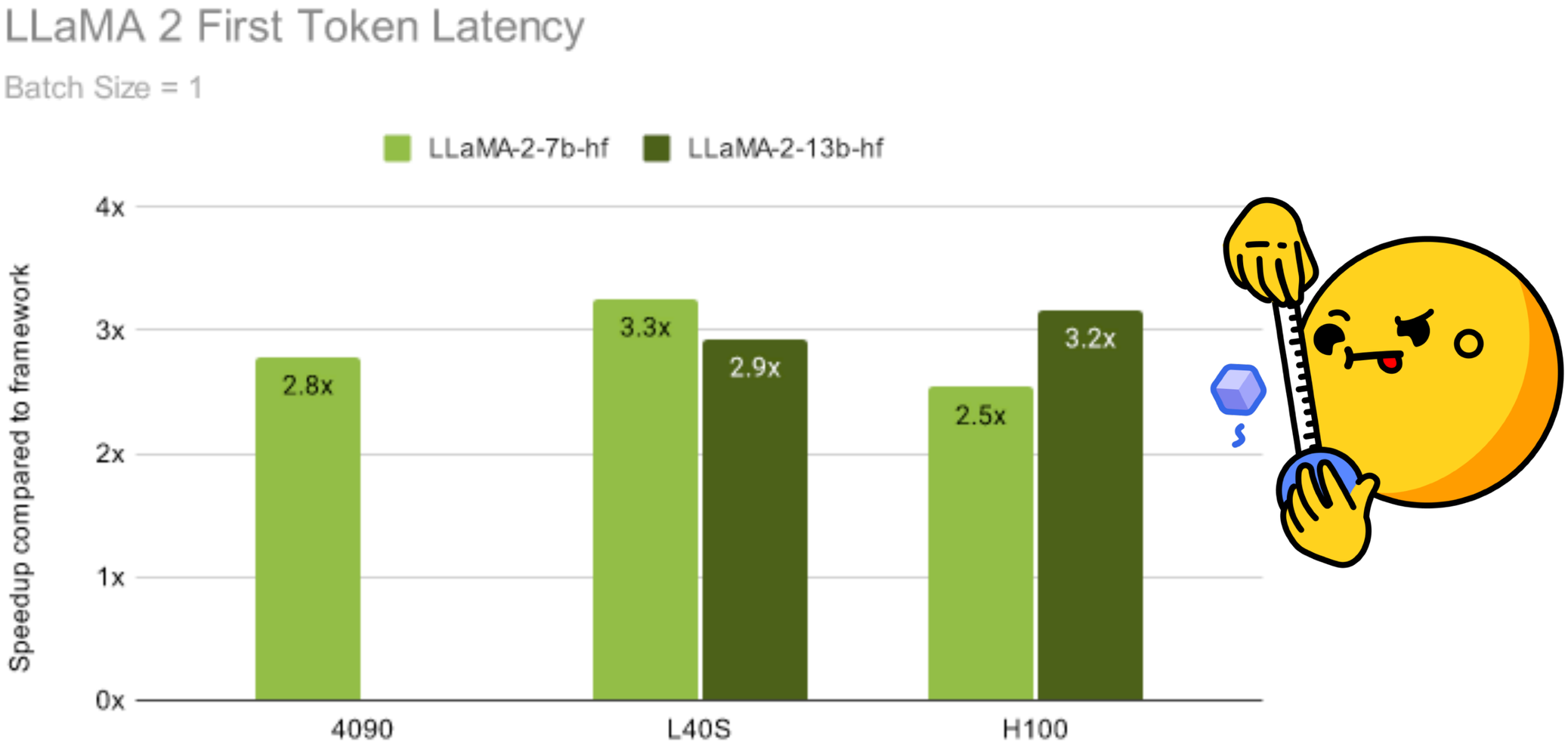
optimum-nvidia

One line code change...

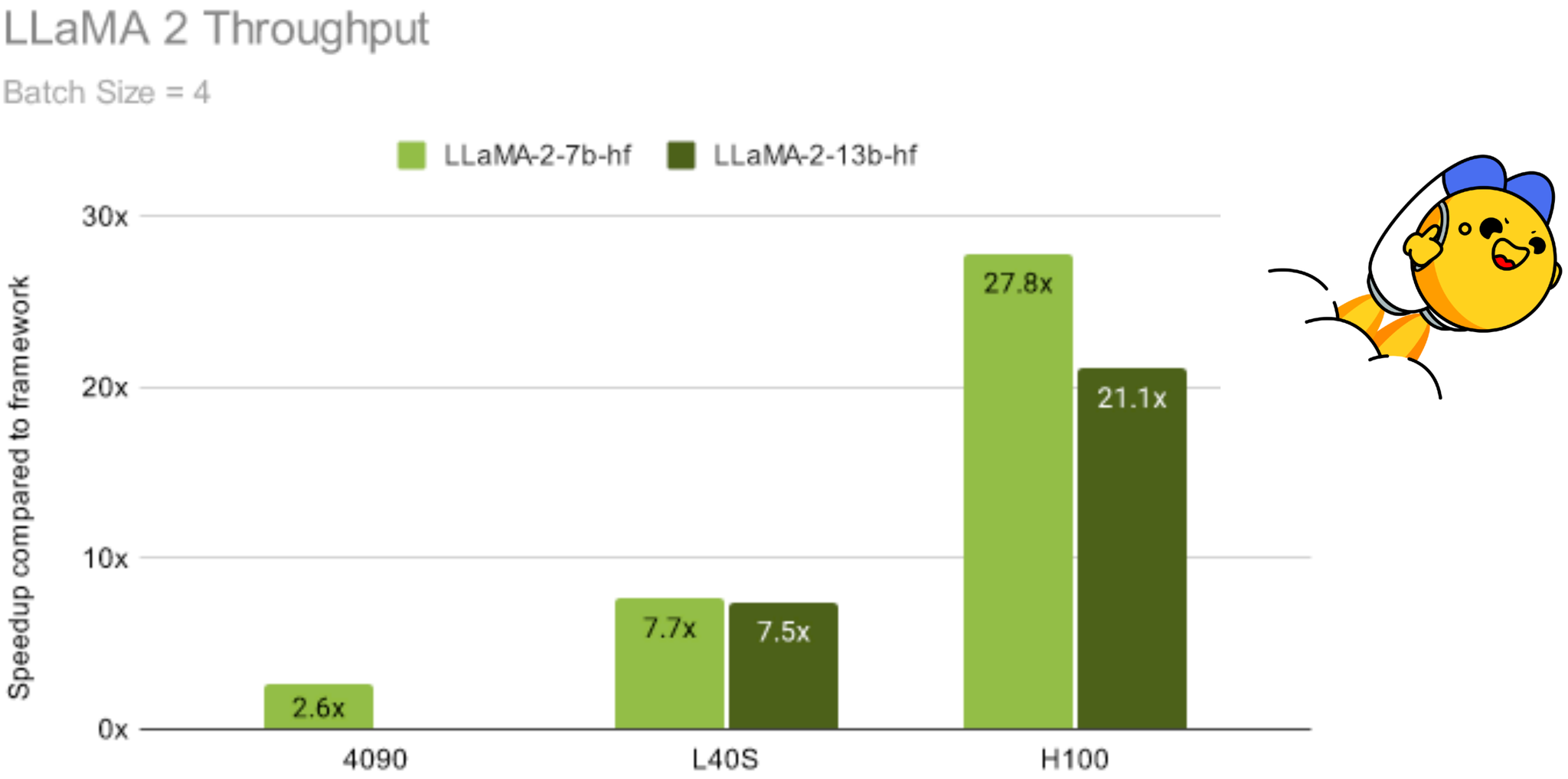
```
- from transformers.pipelines import pipeline
+ from optimum.nvidia.pipelines import pipeline

# 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



And up to 28x throughput!



We're just getting started!

Train on DGX Cloud

optimum-nvidia

NIM Microservices

NeMo Data Curator

>100 models on hf.co/nvidia

Building Accelerated AI with Hugging Face + NVIDIA

GTC March 2024

Jeff Boudier, Product @ 🤗

Thank you!