Introduction to Hugging Face Dataset

AIE 1901 - AI Exploration: LLM for Optimization

Jie Wang, 2025/09/23

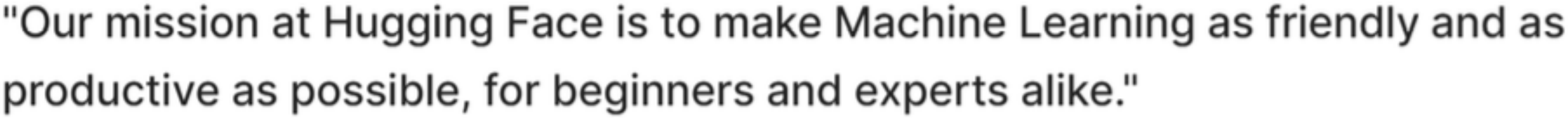
Outline

- Introduction to Hugging Face
- Hands-on Lab: Exploring the Hugging Face Datasets/Models

References:
https://www.freecodecamp.org/news/get-started-with-hugging-face/#heading-how-to-find-the-right-pre-trained-model
https://huggingface.co/docs/datasets/main/en/tutorial
https://huggingface.co/docs/datasets/main/en/quickstart

What is Hugging Face?

- A cute emoji
- A hub where data scientists, researchers, and ML engineers converge to exchange ideas, seek support, and contribute to open-source initiatives
- In their own words, the mission is:



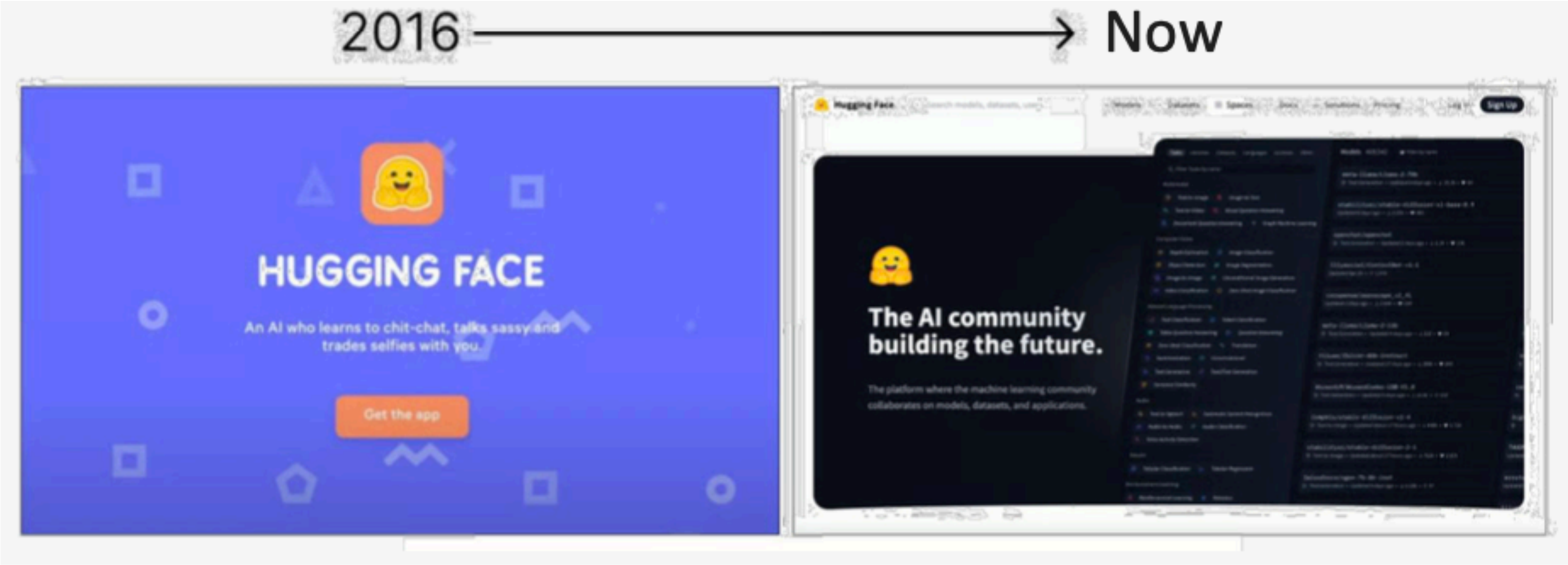
"Our mission at Hugging Face is to make Machine Learning as friendly and as productive as possible, for beginners and experts alike."

What is Hugging Face?

- Providing a simple interface, easy to get started
- Committed to democratizing AI and making it accessible to a global community

What is Hugging Face?

- Founded in 2016 (To develop an interactive AI chatbot targeted at teenagers)
- Moved to open source the AI models since 2018
  - “Transformers” library



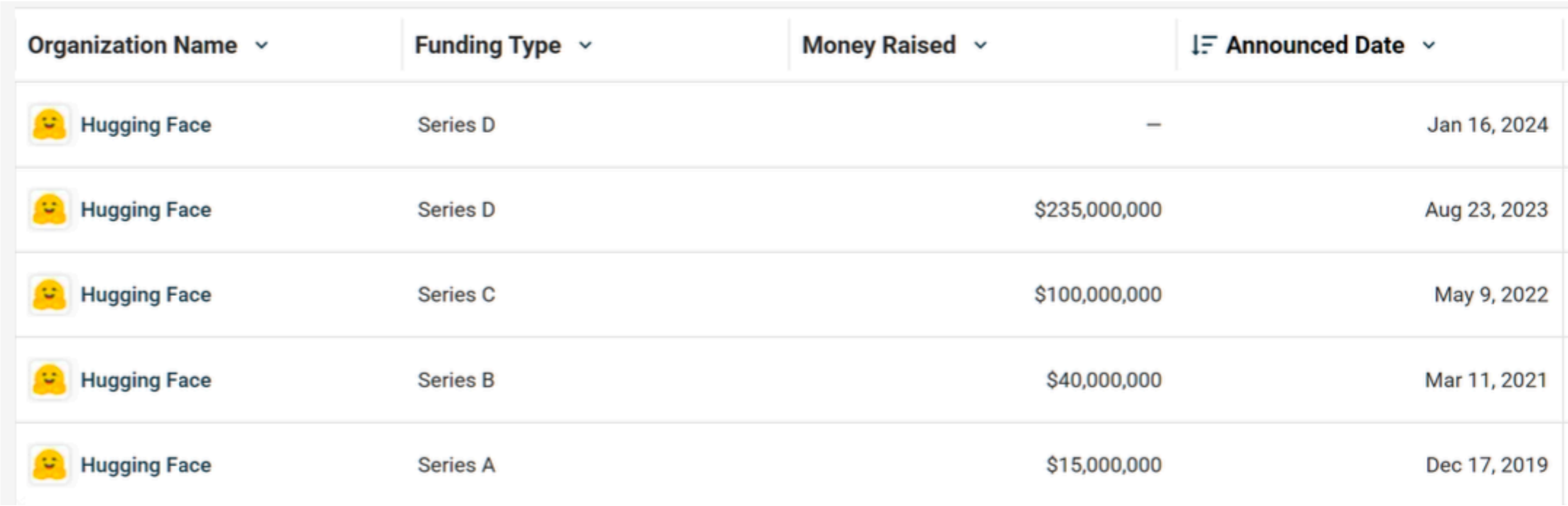
An AI who learns to chit-chat, talks sassy and trades selfies with you.

The AI community building the future.

The platform where the machine learning community collaborates on models, datasets, and applications.

Funding & Valuation

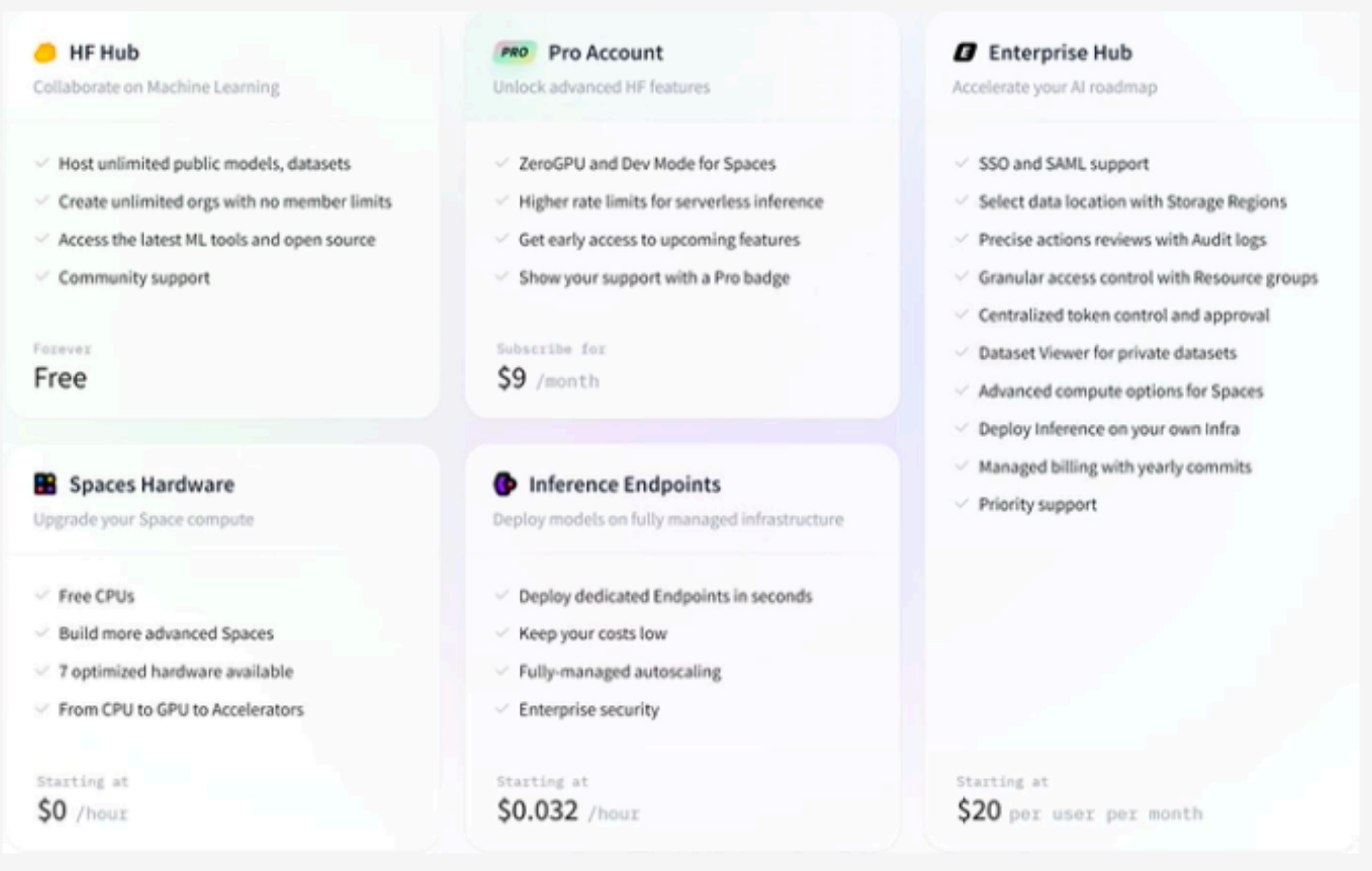
- Hugging Face has a post-money valuation in the range of $1B to $10B as of Aug 24, 2023, according to PrivCo.
- $395.2M in funding over 7 rounds, funded by 38 investors



Organization Name	Funding Type	Money Raised	Announced Date
Hugging Face	Series D	–	Jan 16, 2024
Hugging Face	Series D	$235,000,000	Aug 23, 2023
Hugging Face	Series C	$100,000,000	May 9, 2022
Hugging Face	Series B	$40,000,000	Mar 11, 2021
Hugging Face	Series A	$15,000,000	Dec 17, 2019

Business Models

- Most of Hugging Face's revenue is generated through its **Enterprise** solution.
- Provides a **premium subscription plan** for businesses requiring high-performance AI models and additional features



HF Hub

Collaborate on Machine Learning

- ✓ Host unlimited public models, datasets
- ✓ Create unlimited orgs with no member limits
- ✓ Access the latest ML tools and open source
- ✓ Community support

Forever

Free

PRO Pro Account

Unlock advanced HF features

- ✓ ZeroGPU and Dev Mode for Spaces
- ✓ Higher rate limits for serverless inference
- ✓ Get early access to upcoming features
- ✓ Show your support with a Pro badge

Subscribe for

$9 /month

Enterprise Hub

Accelerate your AI roadmap

- ✓ SSO and SAML support
- ✓ Select data location with Storage Regions
- ✓ Precise actions reviews with Audit logs
- ✓ Granular access control with Resource groups
- ✓ Centralized token control and approval
- ✓ Dataset Viewer for private datasets
- ✓ Advanced compute options for Spaces
- ✓ Deploy Inference on your own Infra
- ✓ Managed billing with yearly commits
- ✓ Priority support

Starting at

$20 per user per month

Spaces Hardware

Upgrade your Space compute

- ✓ Free CPUs
- ✓ Build more advanced Spaces
- ✓ 7 optimized hardware available
- ✓ From CPU to GPU to Accelerators

Starting at

$0 /hour

Inference Endpoints

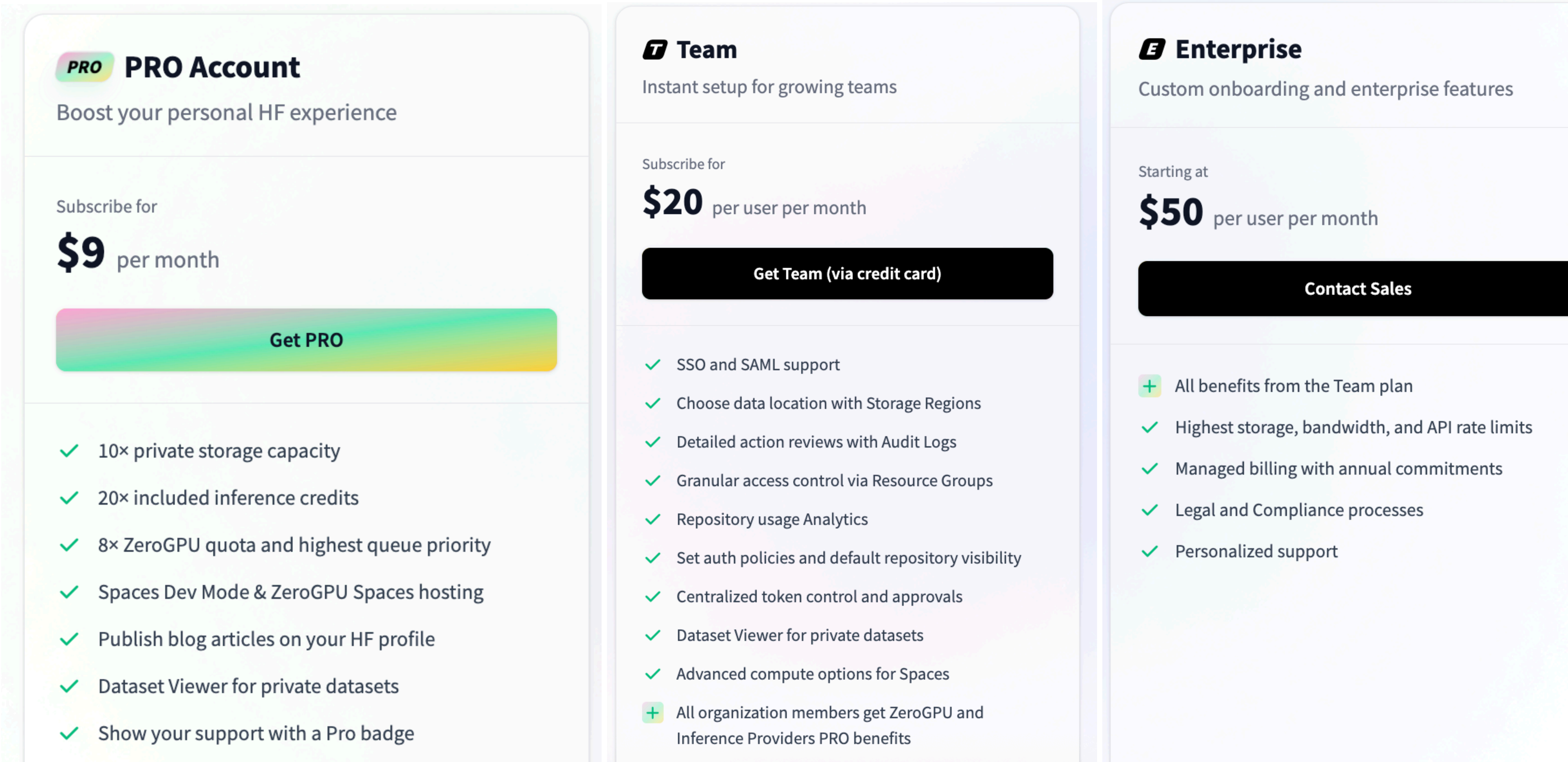
Deploy models on fully managed infrastructure

- ✓ Deploy dedicated Endpoints in seconds
- ✓ Keep your costs low
- ✓ Fully-managed autoscaling
- ✓ Enterprise security

Starting at

$0.032 /hour

Business Models



PRO **PRO Account**

Boost your personal HF experience

Subscribe for

$9 per month

Get PRO

- ✓ 10× private storage capacity
- ✓ 20× included inference credits
- ✓ 8× ZeroGPU quota and highest queue priority
- ✓ Spaces Dev Mode & ZeroGPU Spaces hosting
- ✓ Publish blog articles on your HF profile
- ✓ Dataset Viewer for private datasets
- ✓ Show your support with a Pro badge

Team

Instant setup for growing teams

Subscribe for

$20 per user per month

Get Team (via credit card)

- ✓ SSO and SAML support
- ✓ Choose data location with Storage Regions
- ✓ Detailed action reviews with Audit Logs
- ✓ Granular access control via Resource Groups
- ✓ Repository usage Analytics
- ✓ Set auth policies and default repository visibility
- ✓ Centralized token control and approvals
- ✓ Dataset Viewer for private datasets
- ✓ Advanced compute options for Spaces
- + All organization members get ZeroGPU and Inference Providers PRO benefits

Enterprise

Custom onboarding and enterprise features

Starting at

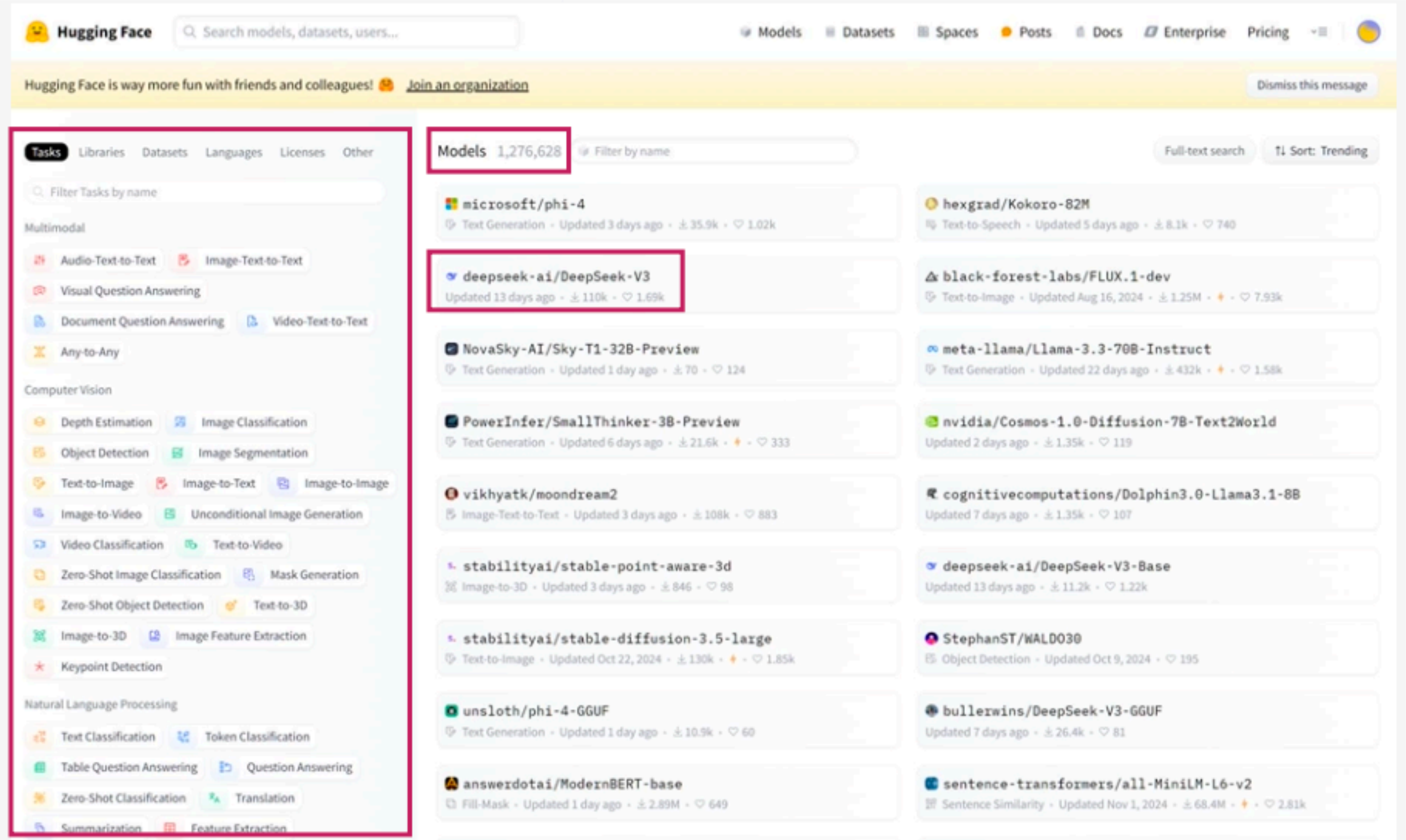
$50 per user per month

Contact Sales

- + All benefits from the Team plan
- ✓ Highest storage, bandwidth, and API rate limits
- ✓ Managed billing with annual commitments
- ✓ Legal and Compliance processes
- ✓ Personalized support

What Can You Do on the Hugging Face Platform?

- Download and fine-tune existing Open-Source models



Hugging Face   Search models, datasets, users...   Models   Datasets   Spaces   Posts   Docs   Enterprise   Pricing

Hugging Face is way more fun with friends and colleagues! Join an organization   Dismiss this message

Multimodal

Audio-Text-to-Text · Image-Text-to-Text · Visual Question Answering · Document Question Answering · Video-Text-to-Text · Any-to-Any

Computer Vision

Depth Estimation · Image Classification · Object Detection · Image Segmentation · Text-to-Image · Image-to-Text · Image-to-Image · Image-to-Video · Unconditional Image Generation · Video Classification · Text-to-Video · Zero-Shot Image Classification · Mask Generation · Zero-Shot Object Detection · Text-to-3D · Image-to-3D · Image Feature Extraction · Keypoint Detection

Natural Language Processing

Text Classification · Token Classification · Table Question Answering · Question Answering · Zero-Shot Classification · Translation · Summarization · Feature Extraction

Models 1,276,628   Filter by name   Full-text search   Sort: Trending

microsoft/phi-4
Text Generation · Updated 3 days ago · ↓ 35.9k · ♡ 1.02k

deepseek-ai/DeepSeek-V3
Updated 13 days ago · ↓ 110k · ♡ 1.69k

NovaSky-AI/Sky-T1-32B-Preview
Text Generation · Updated 1 day ago · ↓ 70 · ♡ 124

PowerInfer/SmallThinker-3B-Preview
Text Generation · Updated 6 days ago · ↓ 21.6k · ♡ 333

vikhyatk/moondream2
Image-Text-to-Text · Updated 3 days ago · ↓ 108k · ♡ 883

stabilityai/stable-point-aware-3d
Image-to-3D · Updated 3 days ago · ↓ 846 · ♡ 98

stabilityai/stable-diffusion-3.5-large
Text-to-Image · Updated Oct 22, 2024 · ↓ 130k · ♡ 1.85k

unsloth/phi-4-GGUF
Text Generation · Updated 1 day ago · ↓ 10.9k · ♡ 60

answerdotai/ModernBERT-base
Fill-Mask · Updated 1 day ago · ↓ 2.89M · ♡ 649

hexgrad/Kokoro-82M
Text-to-Speech · Updated 5 days ago · ↓ 8.1k · ♡ 740

black-forest-labs/FLUX.1-dev
Text-to-Image · Updated Aug 16, 2024 · ↓ 1.25M · ♡ 7.93k

meta-llama/Llama-3.3-70B-Instruct
Text Generation · Updated 22 days ago · ↓ 432k · ♡ 1.58k

nvidia/Cosmos-1.0-Diffusion-7B-Text2World
Updated 2 days ago · ↓ 1.35k · ♡ 119

cognitivecomputations/Dolphin3.0-Llama3.1-8B
Updated 7 days ago · ↓ 1.35k · ♡ 107

deepseek-ai/DeepSeek-V3-Base
Updated 13 days ago · ↓ 11.2k · ♡ 1.22k

StephanST/WALDO30
Object Detection · Updated Oct 9, 2024 · ♡ 195

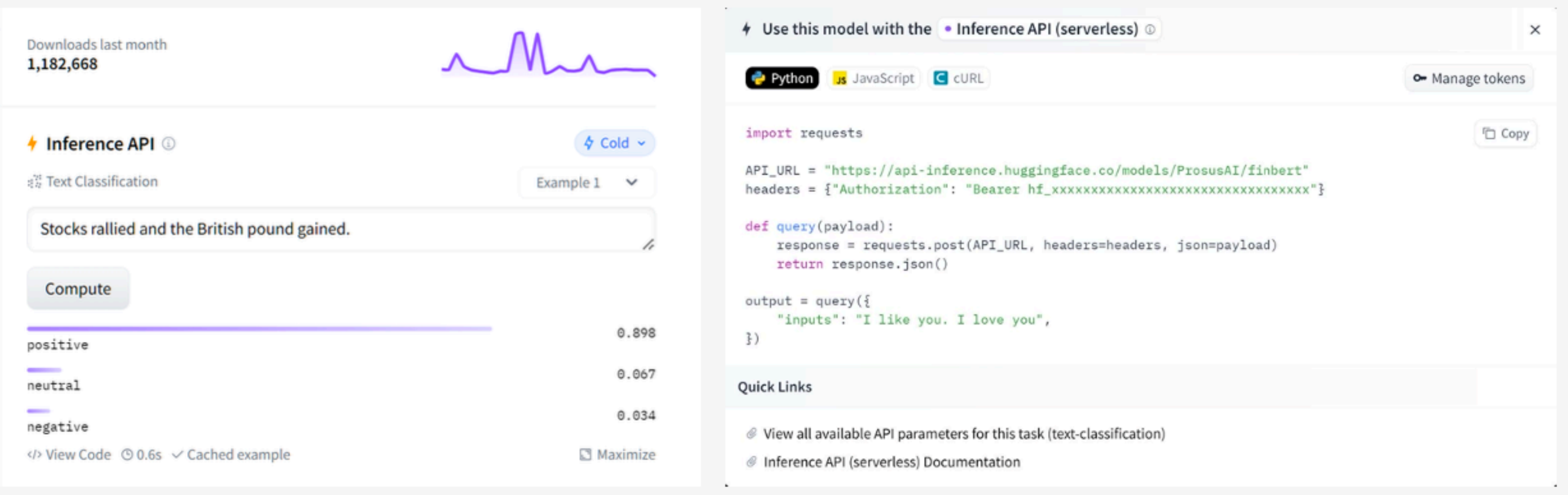
bullerwins/DeepSeek-V3-GGUF
Updated 7 days ago · ↓ 26.4k · ♡ 81

sentence-transformers/all-MiniLM-L6-v2
Sentence Similarity · Updated Nov 1, 2024 · ↓ 68.4M · ♡ 2.81k

What Can You Do on the Hugging Face Platform?

- Run Models directly using Inference API

- If you don’t want to build those models on your own machines, you can simply connect to these models, send requests, and receive outputs via API.



Downloads last month
1,182,668

Inference API

Text Classification

Example 1

Stocks rallied and the British pound gained.

```
positive   0.898
neutral    0.067
negative   0.034
```

Use this model with the Inference API (serverless)

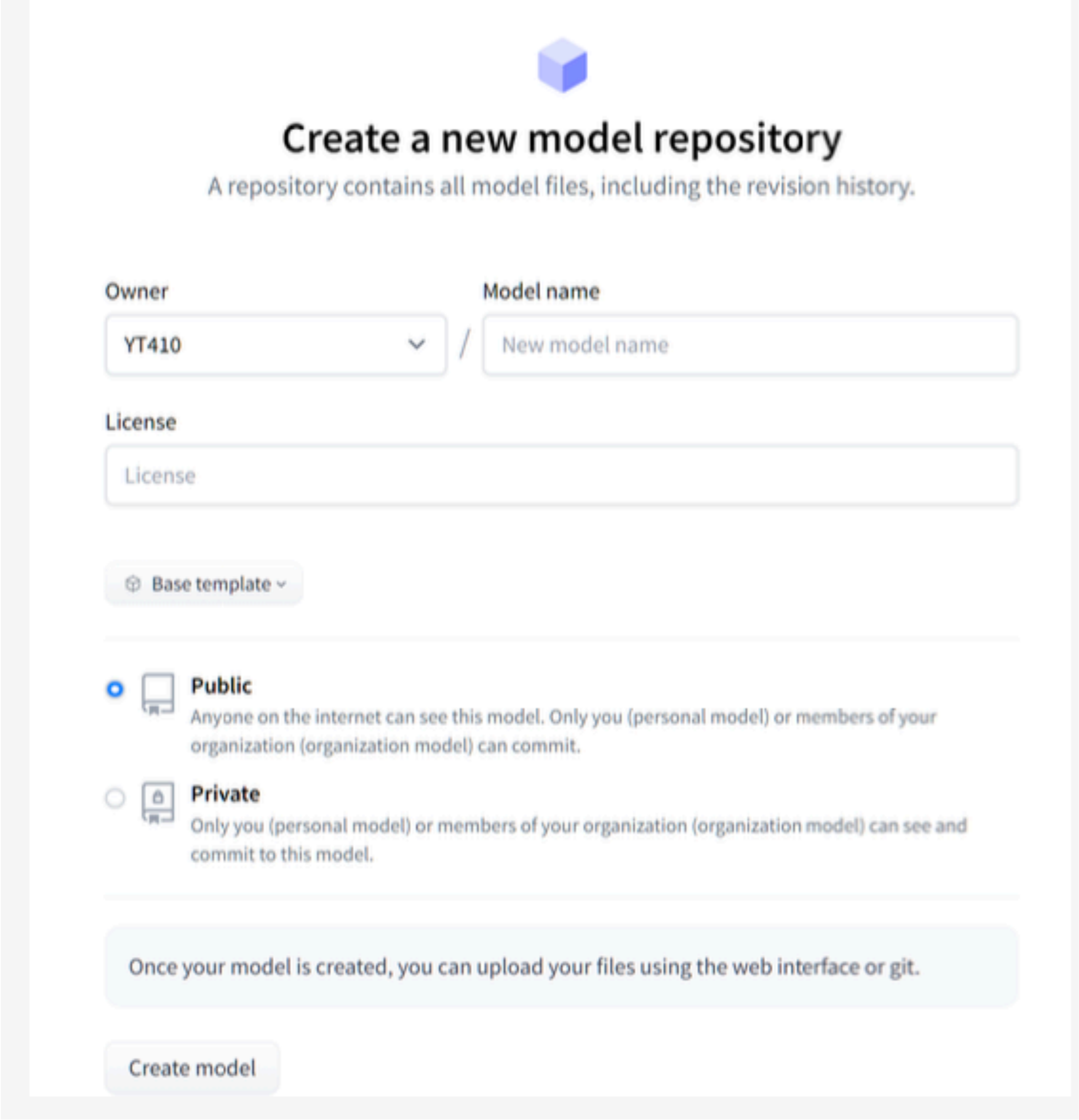
```python
import requests

API_URL = "https://api-inference.huggingface.co/models/ProsusAI/finbert"
headers = {"Authorization": "Bearer hf_xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"}

def query(payload):
    response = requests.post(API_URL, headers=headers, json=payload)
    return response.json()

output = query({
    "inputs": "I like you. I love you",
})
```

Quick Links

- View all available API parameters for this task (text-classification)
- Inference API (serverless) Documentation

Check https://huggingface.co/ProsusAI/finbert

What Can You Do on the Hugging Face Platform?



Create a new model repository

A repository contains all model files, including the revision history.

Owner: YT410

Model name: New model name

License: License

● Public
Anyone on the internet can see this model. Only you (personal model) or members of your organization (organization model) can commit.

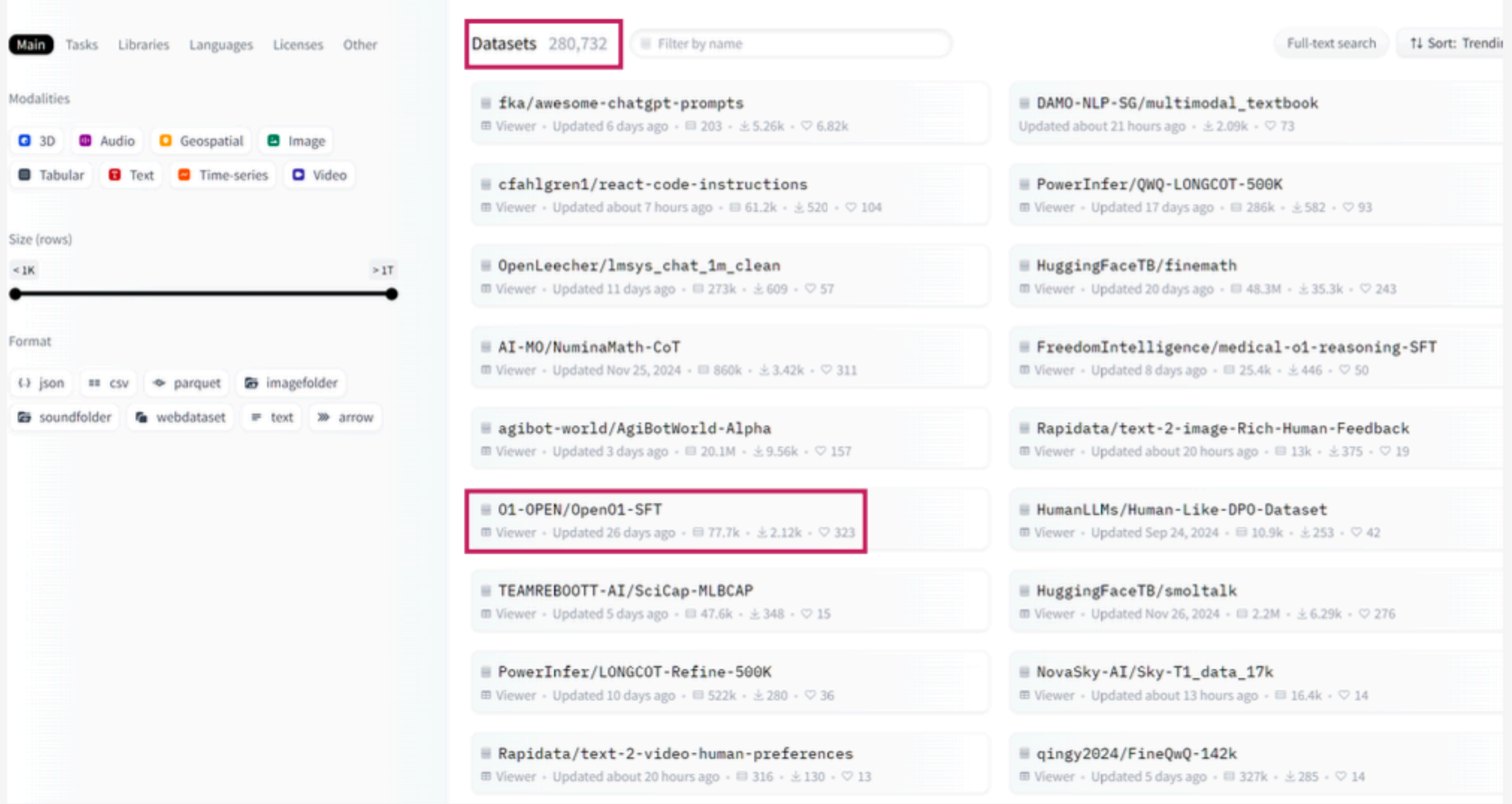
○ Private
Only you (personal model) or members of your organization (organization model) can see and commit to this model.

Once your model is created, you can upload your files using the web interface or git.

- Add/create your own model
- Host your model on the platform and make it public or private

What Can You Do on the Hugging Face Platform?

- Use existing datasets



Datasets 280,732

fka/awesome-chatgpt-prompts
Viewer · Updated 6 days ago · 203 · 5.26k · 6.82k

cfahlgren1/react-code-instructions
Viewer · Updated about 7 hours ago · 61.2k · 520 · 104

OpenLeecher/lmsys_chat_1m_clean
Viewer · Updated 11 days ago · 273k · 609 · 57

AI-MO/NuminaMath-CoT
Viewer · Updated Nov 25, 2024 · 860k · 3.42k · 311

agibot-world/AgiBotWorld-Alpha
Viewer · Updated 3 days ago · 20.1M · 9.56k · 157

O1-OPEN/OpenO1-SFT
Viewer · Updated 26 days ago · 77.7k · 2.12k · 323

TEAMREBOOTT-AI/SciCap-MLBCAP
Viewer · Updated 5 days ago · 47.6k · 348 · 15

PowerInfer/LONGCOT-Refine-500K
Viewer · Updated 10 days ago · 522k · 280 · 36

Rapidata/text-2-video-human-preferences
Viewer · Updated about 20 hours ago · 316 · 130 · 13

DAMO-NLP-SG/multimodal_textbook
Updated about 21 hours ago · 2.09k · 73

PowerInfer/QWQ-LONGCOT-500K
Viewer · Updated 17 days ago · 286k · 582 · 93

HuggingFaceTB/finemath
Viewer · Updated 20 days ago · 48.3M · 35.3k · 243

FreedomIntelligence/medical-o1-reasoning-SFT
Viewer · Updated 8 days ago · 25.4k · 446 · 50

Rapidata/text-2-image-Rich-Human-Feedback
Viewer · Updated about 20 hours ago · 13k · 375 · 19

HumanLLMs/Human-Like-DPO-Dataset
Viewer · Updated Sep 24, 2024 · 10.9k · 253 · 42

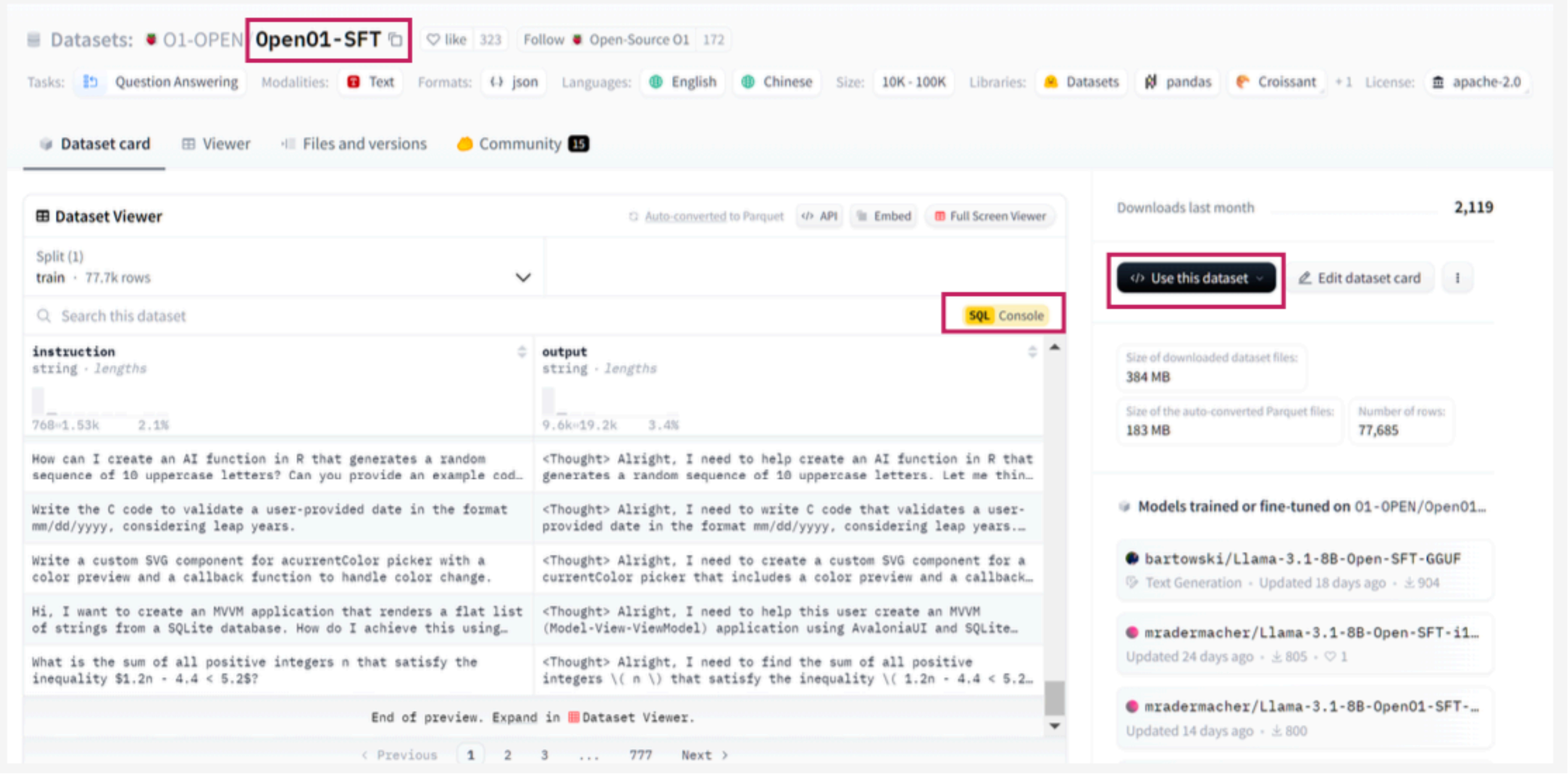
HuggingFaceTB/smoltalk
Viewer · Updated Nov 26, 2024 · 2.2M · 6.29k · 276

NovaSky-AI/Sky-T1_data_17k
Viewer · Updated about 13 hours ago · 16.4k · 14

qingy2024/FineQwQ-142k
Viewer · Updated 5 days ago · 327k · 285 · 14

What Can You Do on the Hugging Face Platform?

- Use existing datasets



Datasets: O1-OPEN/Open01-SFT

like 323 · Follow Open-Source O1 172

Tasks: Question Answering · Modalities: Text · Formats: json · Languages: English, Chinese · Size: 10K - 100K · Libraries: Datasets, pandas, Croissant +1 · License: apache-2.0

Split (1)
train · 77.7k rows

instruction string · lengths 768⇔1.53k 2.1%	output string · lengths 9.6k⇔19.2k 3.4%
How can I create an AI function in R that generates a random sequence of 10 uppercase letters? Can you provide an example cod…	<Thought> Alright, I need to help create an AI function in R that generates a random sequence of 10 uppercase letters. Let me thin…
Write the C code to validate a user-provided date in the format mm/dd/yyyy, considering leap years.	<Thought> Alright, I need to write C code that validates a user-provided date in the format mm/dd/yyyy, considering leap years.…
Write a custom SVG component for acurrentColor picker with a color preview and a callback function to handle color change.	<Thought> Alright, I need to create a custom SVG component for a currentColor picker that includes a color preview and a callback…
Hi, I want to create an MVVM application that renders a flat list of strings from a SQLite database. How do I achieve this using…	<Thought> Alright, I need to help this user create an MVVM (Model-View-ViewModel) application using AvaloniaUI and SQLite…
What is the sum of all positive integers n that satisfy the inequality $1.2n - 4.4 < 5.2$?	<Thought> Alright, I need to find the sum of all positive integers \(n \) that satisfy the inequality \(1.2n - 4.4 < 5.2…

End of preview. Expand in Dataset Viewer.

‹ Previous 1 2 3 ... 777 Next ›

Downloads last month 2,119

Size of downloaded dataset files: 384 MB

Size of the auto-converted Parquet files: 183 MB

Number of rows: 77,685

Models trained or fine-tuned on 01-OPEN/Open01…

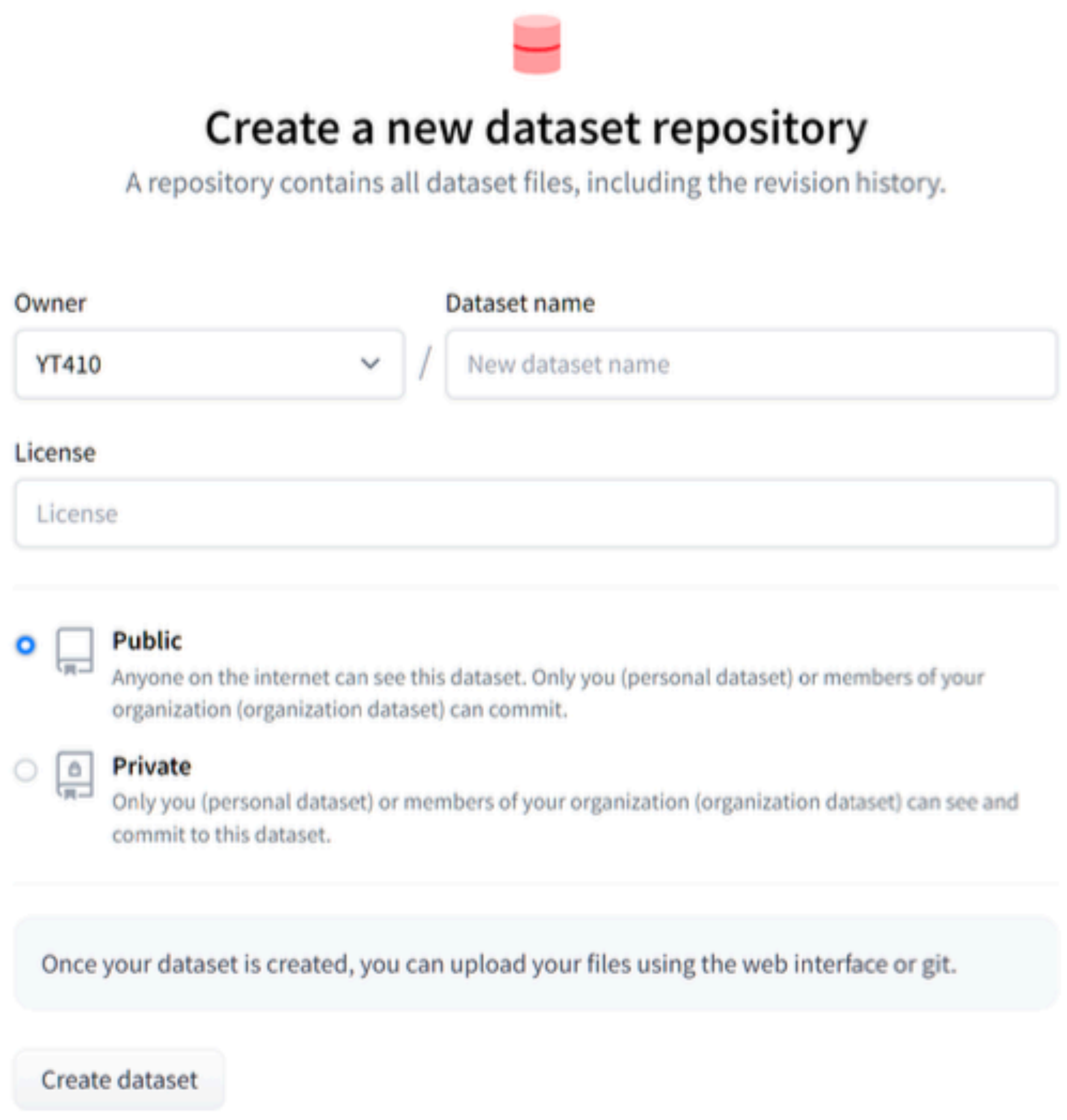
bartowski/Llama-3.1-8B-Open-SFT-GGUF
Text Generation · Updated 18 days ago · 904

mradermacher/Llama-3.1-8B-Open-SFT-i1…
Updated 24 days ago · 805 · 1

mradermacher/Llama-3.1-8B-Open01-SFT-…
Updated 14 days ago · 800

What Can You Do on the Hugging Face Platform?

- Upload your own dataset
- Host your own dataset on the platform and make it public or private



Create a new dataset repository

A repository contains all dataset files, including the revision history.

Owner: YT410 / Dataset name: New dataset name

License: License

● Public
Anyone on the internet can see this dataset. Only you (personal dataset) or members of your organization (organization dataset) can commit.

○ Private
Only you (personal dataset) or members of your organization (organization dataset) can see and commit to this dataset.

Once your dataset is created, you can upload your files using the web interface or git.

What Can You Do on the Hugging Face Platform?

- Create/browse/try out demo apps in Spaces
- Here are a few cool spaces you can check out:
  - OpenAI's Whisper: Transcribe long-form microphone or audio inputs with the click of a button
  - AI Comic Factory: Create your own comic books.
  - QR Code AI Art Generator: Generate beautiful QR codes using AI.
  - Stable Video Diffusion (Img2Vid - XT): Generate 4s video from a single image.
  - Video-LLaMA: Audio-Visual Language Model for Video Understanding.

What Can You Do on the Hugging Face Platform?

- Build yourself as an AI professional

1. **Join or create an organization**

   - You can join or create your own organization on Hugging Face. This allows you to showcase your work and collaborate with other members from your university, lab, or company.

2. **Create a portfolio**

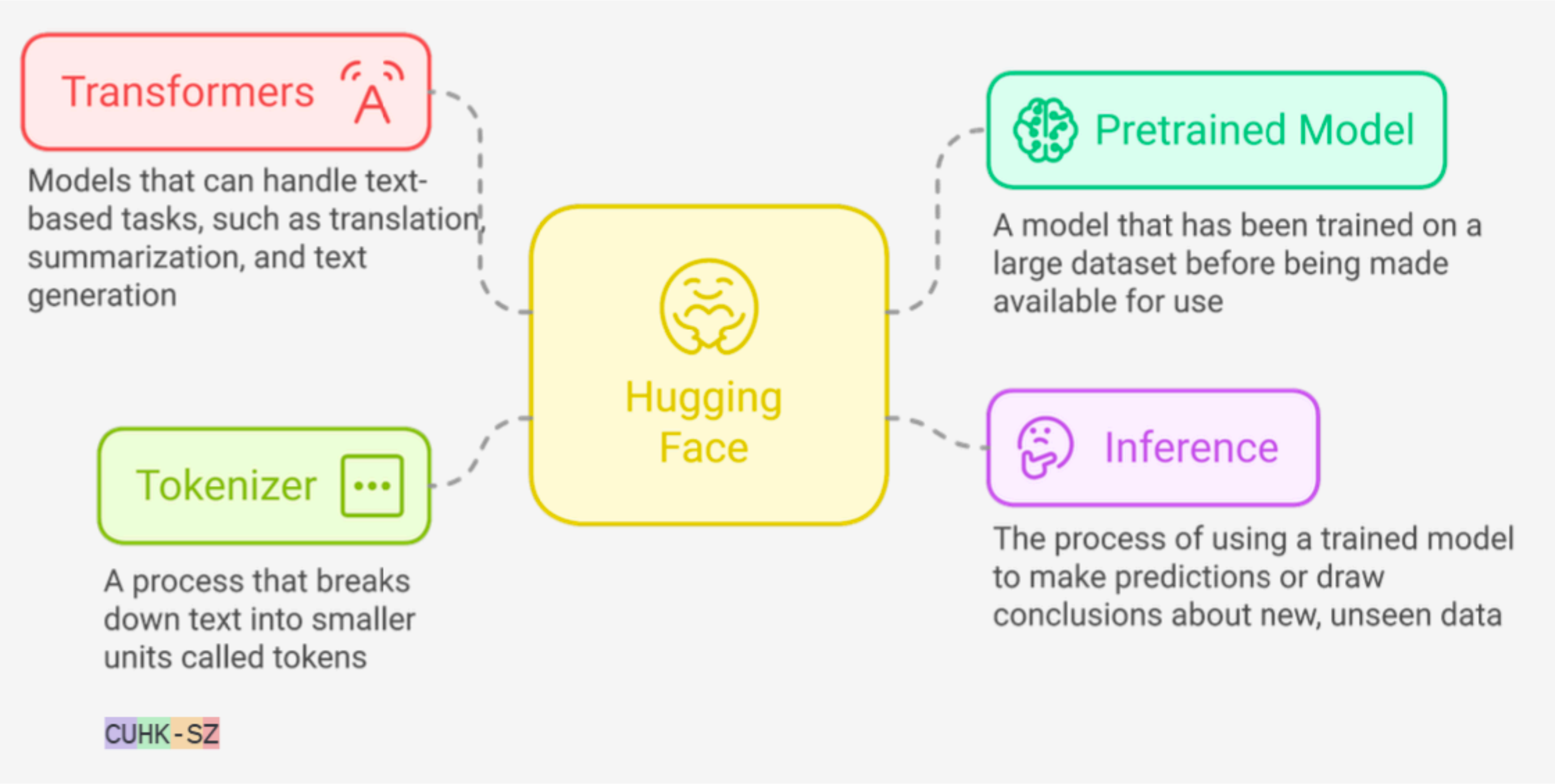
   - You can create a professional portfolio on Hugging Face to showcase your work and start building your reputation. This can help you land jobs related to AI model training, integration, and development.

3. **Learn AI Skills**

   - Hugging Face is an excellent platform for learning AI skills. It offers a comprehensive set of tools and resources for training and using models

   - You can also learn from the experts and the community on Hugging Face, and improve your AI knowledge and skills

Hugging Face Terminology



Hugging Face

Transformers

Models that can handle text-based tasks, such as translation, summarization, and text generation

Pretrained Model

A model that has been trained on a large dataset before being made available for use

Tokenizer

A process that breaks down text into smaller units called tokens

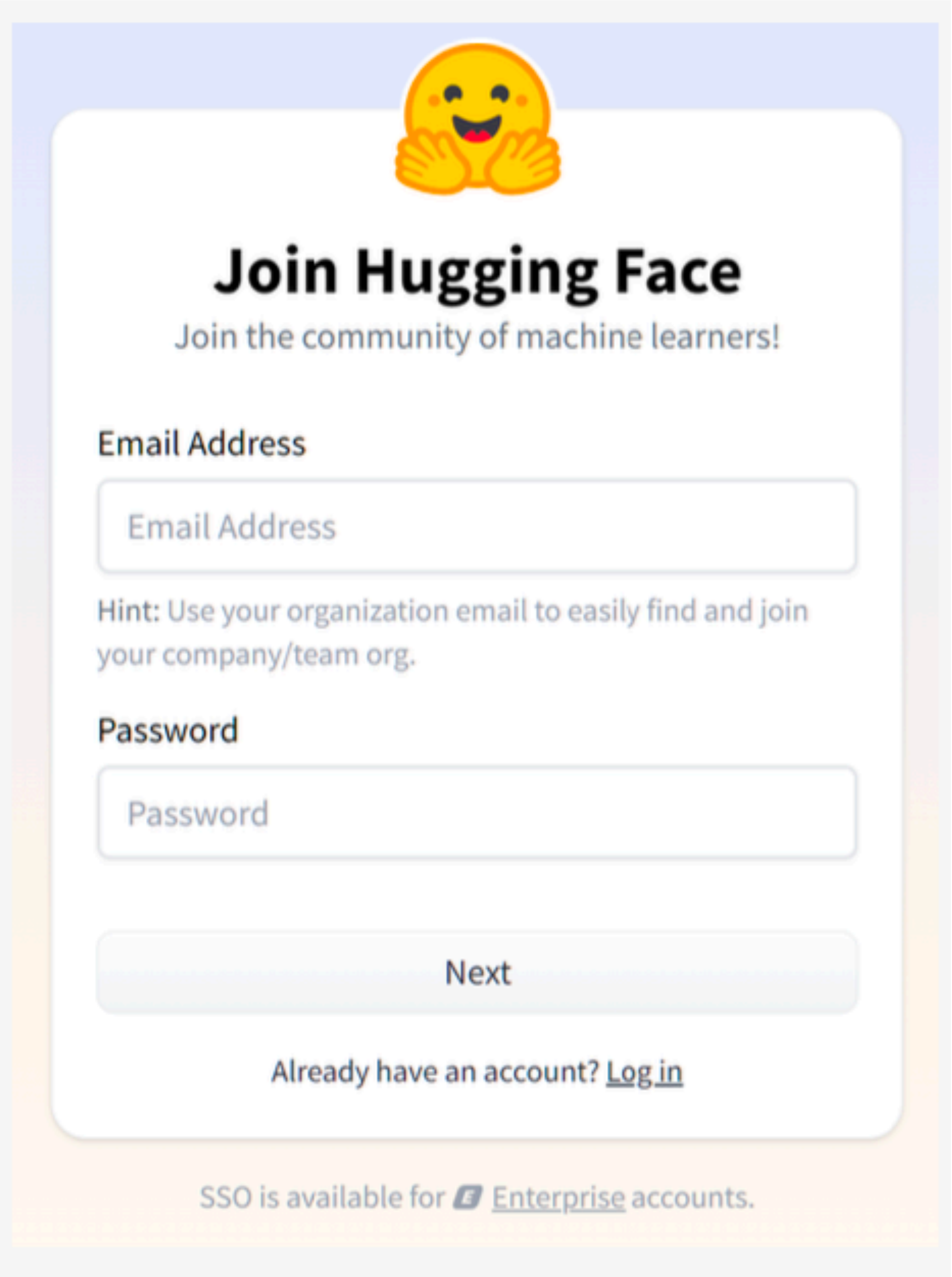
Inference

The process of using a trained model to make predictions or draw conclusions about new, unseen data

CUHK-SZ

How to Get Started with Hugging Face

1. Connect to the VPN
2. Create a Hugging Face Account



Join Hugging Face

Join the community of machine learners!

Email Address

Email Address

Hint: Use your organization email to easily find and join your company/team org.

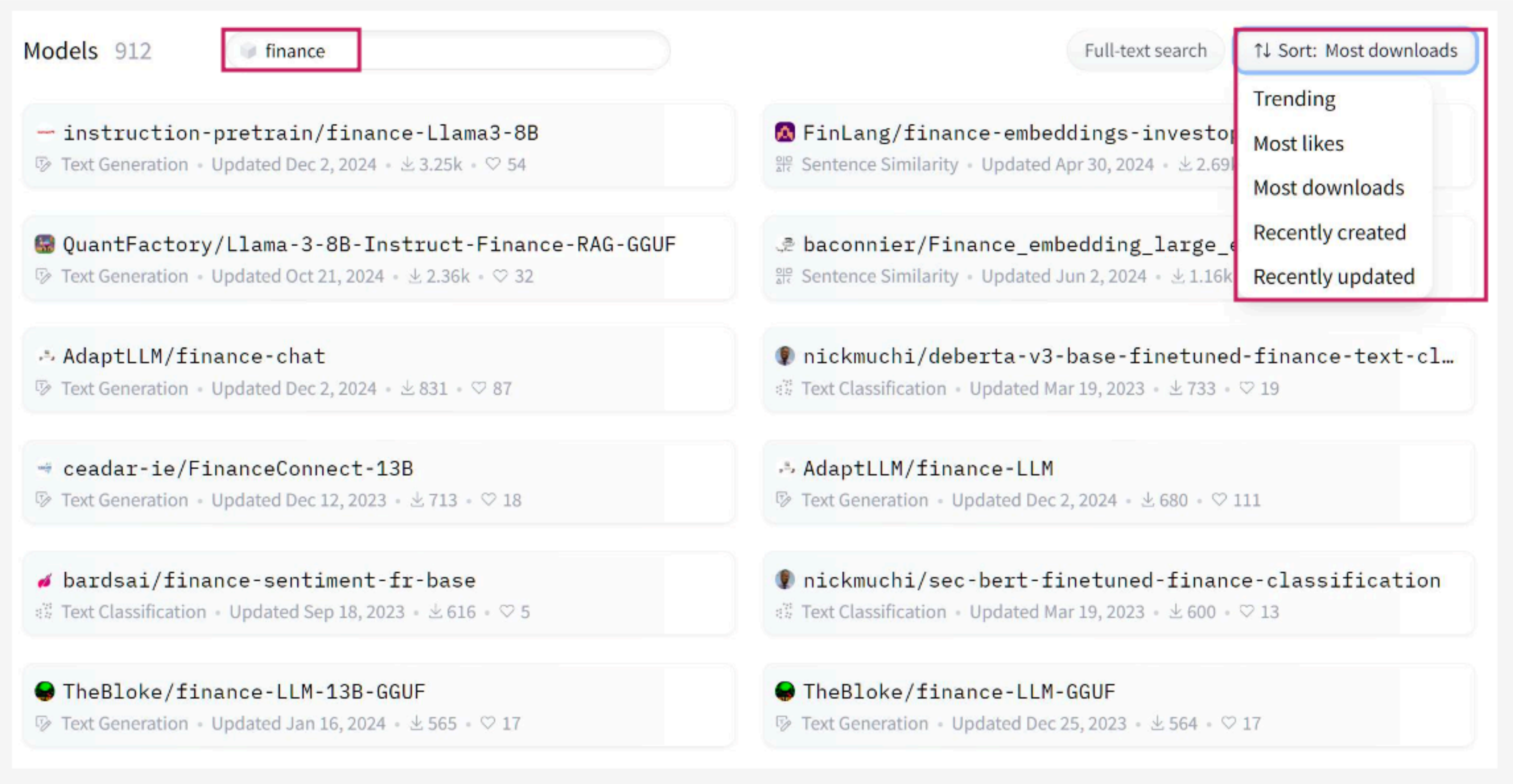
Password

Password

Already have an account? Log in

SSO is available for Enterprise accounts.

How to Find the Right Pre-Trained Model



Models 912

finance

Full-text search

Sort: Most downloads

instruction-pretrain/finance-Llama3-8B
Text Generation • Updated Dec 2, 2024 • 3.25k • 54

FinLang/finance-embeddings-investo
Sentence Similarity • Updated Apr 30, 2024 • 2.69

QuantFactory/Llama-3-8B-Instruct-Finance-RAG-GGUF
Text Generation • Updated Oct 21, 2024 • 2.36k • 32

baconnier/Finance_embedding_large_e
Sentence Similarity • Updated Jun 2, 2024 • 1.16k

AdaptLLM/finance-chat
Text Generation • Updated Dec 2, 2024 • 831 • 87

nickmuchi/deberta-v3-base-finetuned-finance-text-cl…
Text Classification • Updated Mar 19, 2023 • 733 • 19

ceadar-ie/FinanceConnect-13B
Text Generation • Updated Dec 12, 2023 • 713 • 18

AdaptLLM/finance-LLM
Text Generation • Updated Dec 2, 2024 • 680 • 111

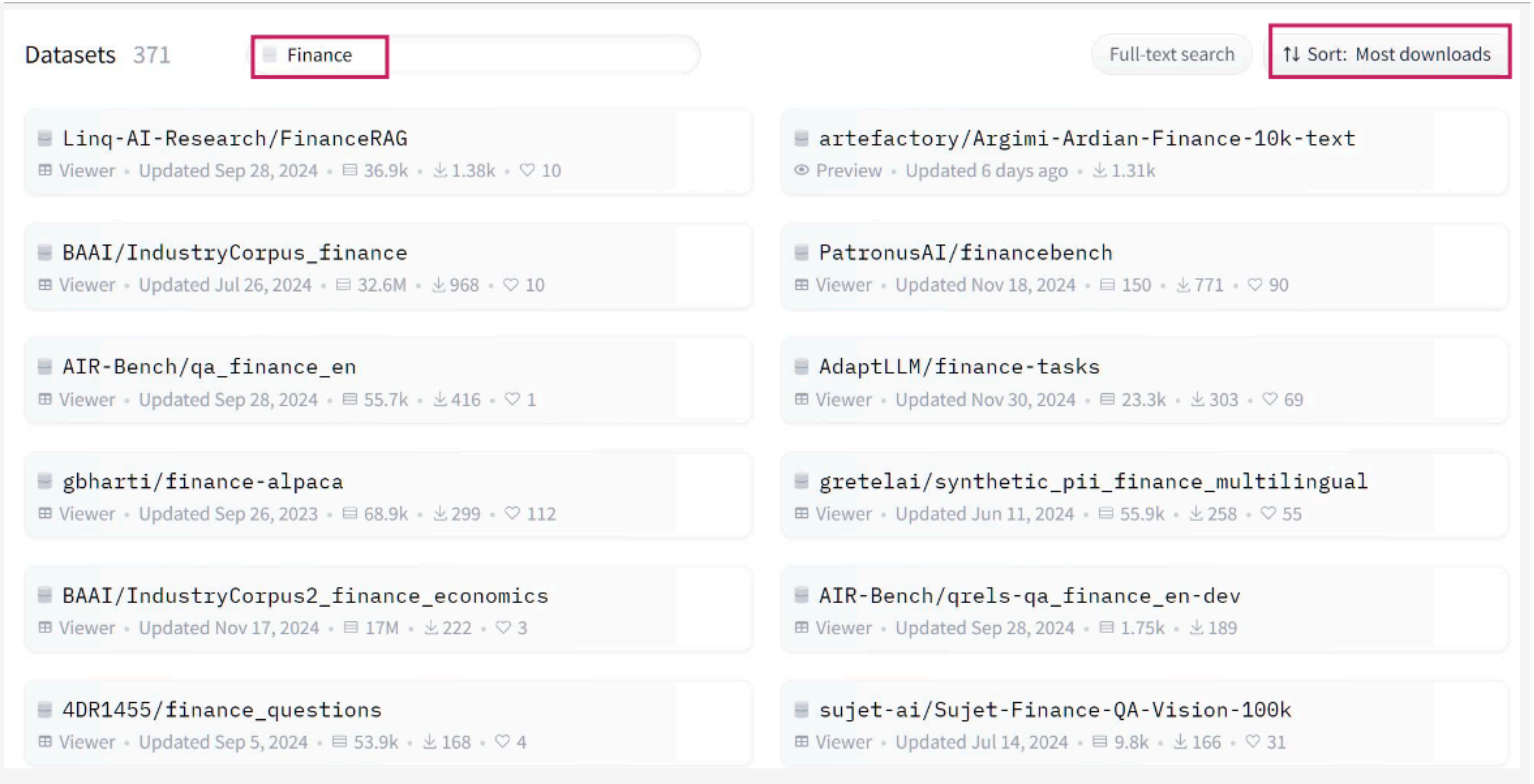
bardsai/finance-sentiment-fr-base
Text Classification • Updated Sep 18, 2023 • 616 • 5

nickmuchi/sec-bert-finetuned-finance-classification
Text Classification • Updated Mar 19, 2023 • 600 • 13

TheBloke/finance-LLM-13B-GGUF
Text Generation • Updated Jan 16, 2024 • 565 • 17

TheBloke/finance-LLM-GGUF
Text Generation • Updated Dec 25, 2023 • 564 • 17

How to Find the Right Dataset



Datasets 371 | Finance | Full-text search | ⇅ Sort: Most downloads

Linq-AI-Research/FinanceRAG
⊞ Viewer • Updated Sep 28, 2024 • ⊟ 36.9k • ↓ 1.38k • ♡ 10

artefactory/Argimi-Ardian-Finance-10k-text
Preview • Updated 6 days ago • ↓ 1.31k

BAAI/IndustryCorpus_finance
⊞ Viewer • Updated Jul 26, 2024 • ⊟ 32.6M • ↓ 968 • ♡ 10

PatronusAI/financebench
⊞ Viewer • Updated Nov 18, 2024 • ⊟ 150 • ↓ 771 • ♡ 90

AIR-Bench/qa_finance_en
⊞ Viewer • Updated Sep 28, 2024 • ⊟ 55.7k • ↓ 416 • ♡ 1

AdaptLLM/finance-tasks
⊞ Viewer • Updated Nov 30, 2024 • ⊟ 23.3k • ↓ 303 • ♡ 69

gbharti/finance-alpaca
⊞ Viewer • Updated Sep 26, 2023 • ⊟ 68.9k • ↓ 299 • ♡ 112

gretelai/synthetic_pii_finance_multilingual
⊞ Viewer • Updated Jun 11, 2024 • ⊟ 55.9k • ↓ 258 • ♡ 55

BAAI/IndustryCorpus2_finance_economics
⊞ Viewer • Updated Nov 17, 2024 • ⊟ 17M • ↓ 222 • ♡ 3

AIR-Bench/qrels-qa_finance_en-dev
⊞ Viewer • Updated Sep 28, 2024 • ⊟ 1.75k • ↓ 189

4DR1455/finance_questions
⊞ Viewer • Updated Sep 5, 2024 • ⊟ 53.9k • ↓ 168 • ♡ 4

sujet-ai/Sujet-Finance-QA-Vision-100k
⊞ Viewer • Updated Jul 14, 2024 • ⊟ 9.8k • ↓ 166 • ♡ 31

Explore LLM Datasets in Hugging Face

- **CardinalOperations/OR-Instruct-Data-3K**

○ Use Case: Instruction dataset to train large language models to learn optimization modeling and solving

○ Scope: **Applications from various industrial problems**

○ Size: 3000

https://huggingface.co/datasets/CardinalOperations/OR-Instruct-Data-3K

Explore LLM Datasets in Hugging Face

- **CardinalOperations/OR-Instruct-Data-3K**

○ Use Case: Instruction dataset to train large language models to learn optimization modeling and solving

○ Scope: **Applications from various industrial problems**

○ Size: 3000

https://huggingface.co/datasets/CardinalOperations/OR-Instruct-Data-3K

- **CardinalOperations/IndustryOR**

○ Use Case: Evaluate LLM performance on optimization.

○ Scope: **a large-scale collection of industrial problems**

○ Size: 100

○ https://huggingface.co/datasets/CardinalOperations/IndustryOR

Hands-on Lab1: Create your Google Account and Log Into Google Colab

- Please refer to 注册谷歌账号教程.pdf

Hands-on Lab2: Try LLM Platform

- Open website https://cloud.siliconflow.cn/playground/chat/
- Create your own dataset
- Example Prompt: “give me 10 sentences that end with the word bird”
- Try it on different LLMs such as Owen2.5-7B-Instruct, Kimi-K2-Instruct-0905, DeepSeek-R1, DeepSeek-V3, Ling-flash 2.0

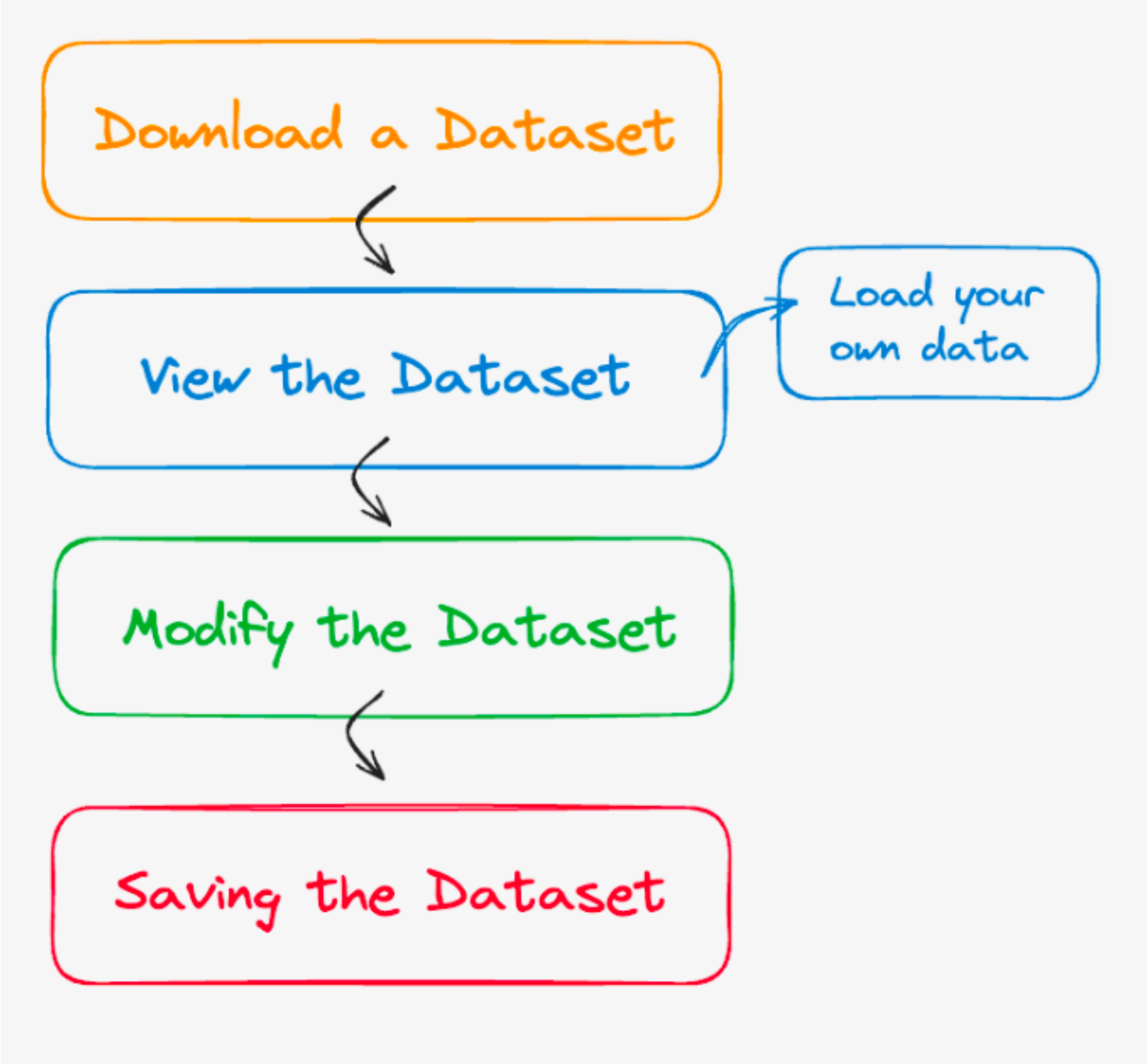
Hands-on Lab3: Exploring the Hugging Face Datasets/Models

- Open the Colab Notebook: https://colab.research.google.com/drive/11tPb_t2KPnzCYTZpSleY9BffbM6U81p9?usp=sharing

- To work on this notebook, go to **File → Save a copy in Drive**

- Work on your copy, leaving this original unchanged

- All your work will be saved in your Google Drive



Download a Dataset
View the Dataset
Load your own data
Modify the Dataset
Saving the Dataset