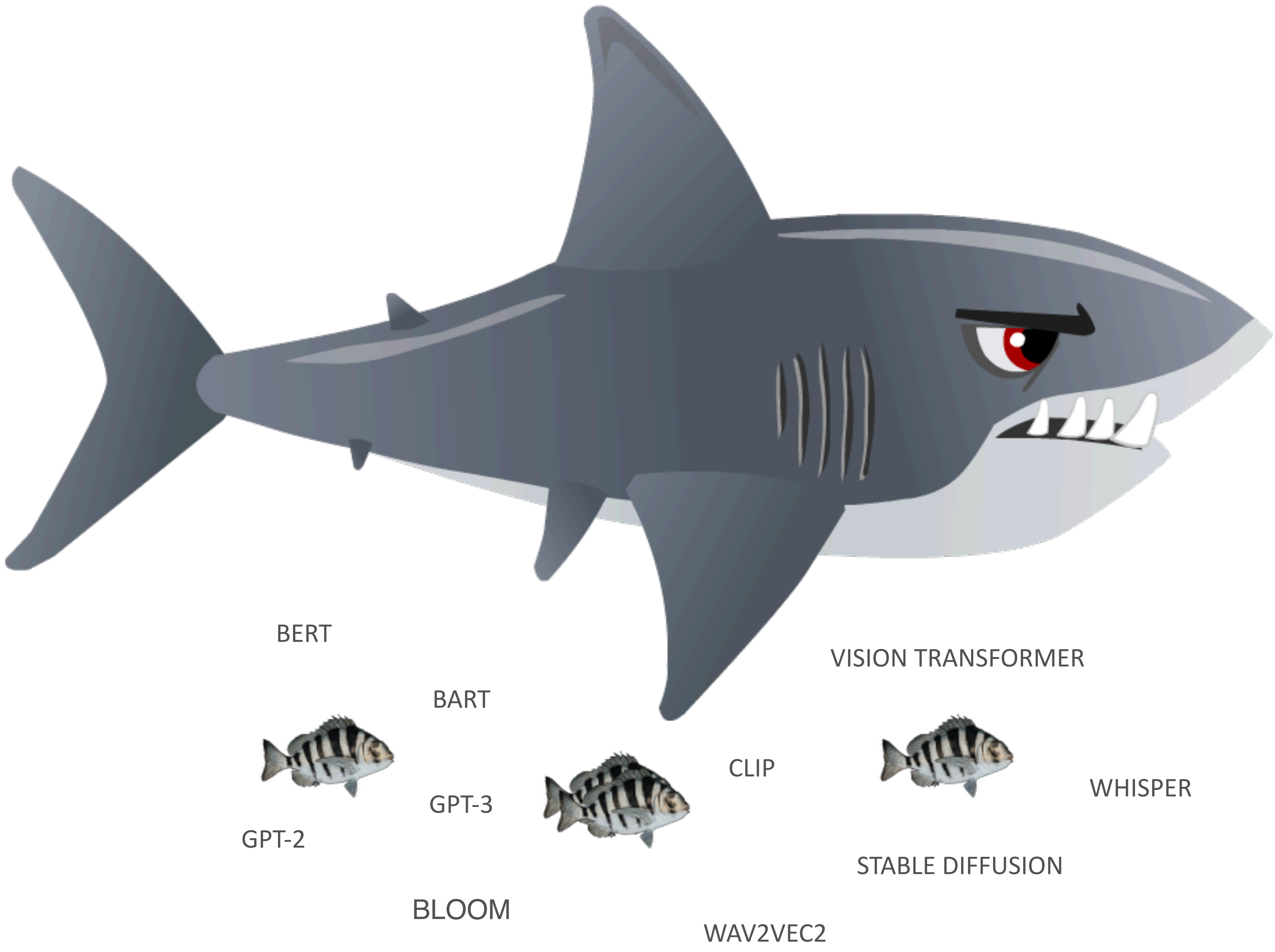
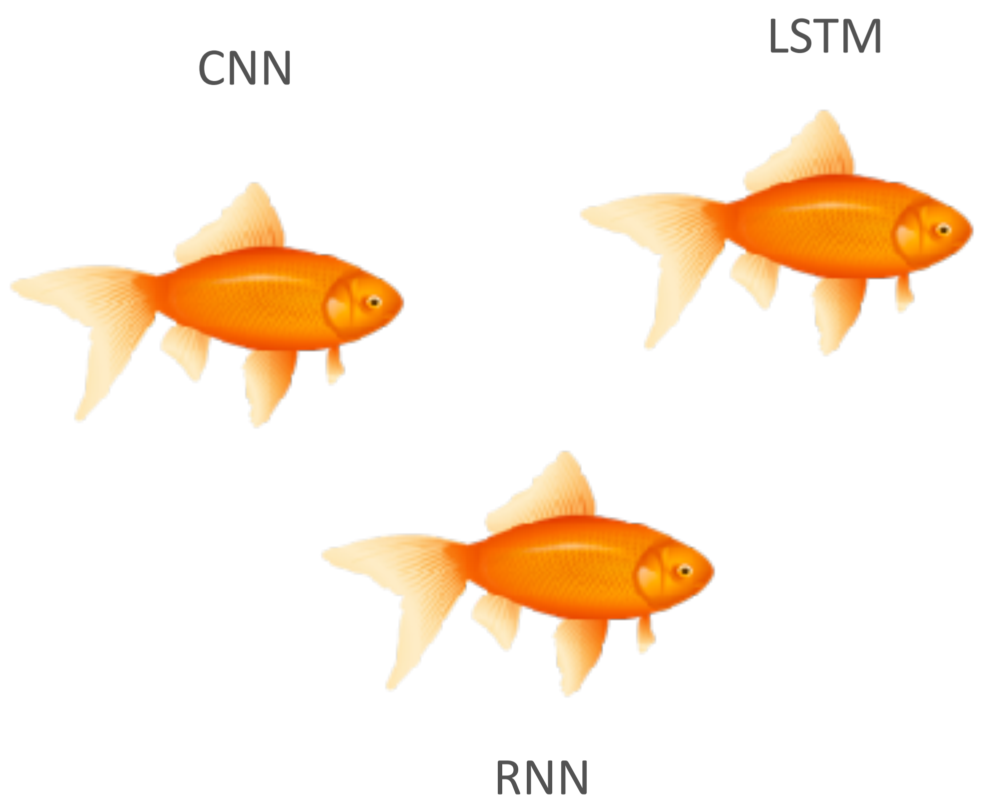
# Demystifying the technology behind Generative AI

Julien Simon, Chief Evangelist, Hugging Face
julsimon@huggingface.co

# 2022: Transformers are eating Deep Learning

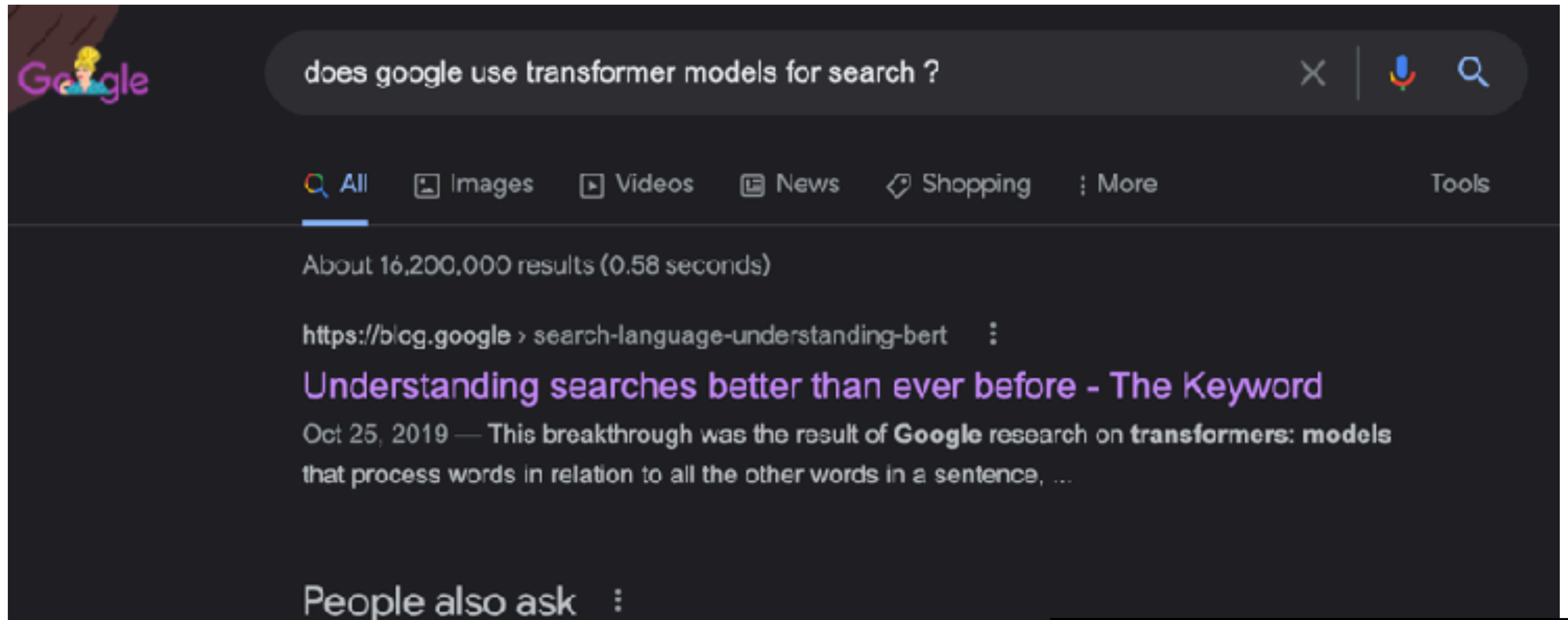




"*Transformers are emerging as*
*a general-purpose architecture for ML*"
[https://www.stateof.ai](https://www.stateof.ai) (2021)

RNN and CNN usage down, Transformers usage up!
[https://www.kaggle.com/kaggle-survey-2021](https://www.kaggle.com/kaggle-survey-2021)

# Transformer models in the wild



does google use transformer models for search ?

About 16,200,000 results (0.58 seconds)

https://blog.google › search-language-understanding-bert

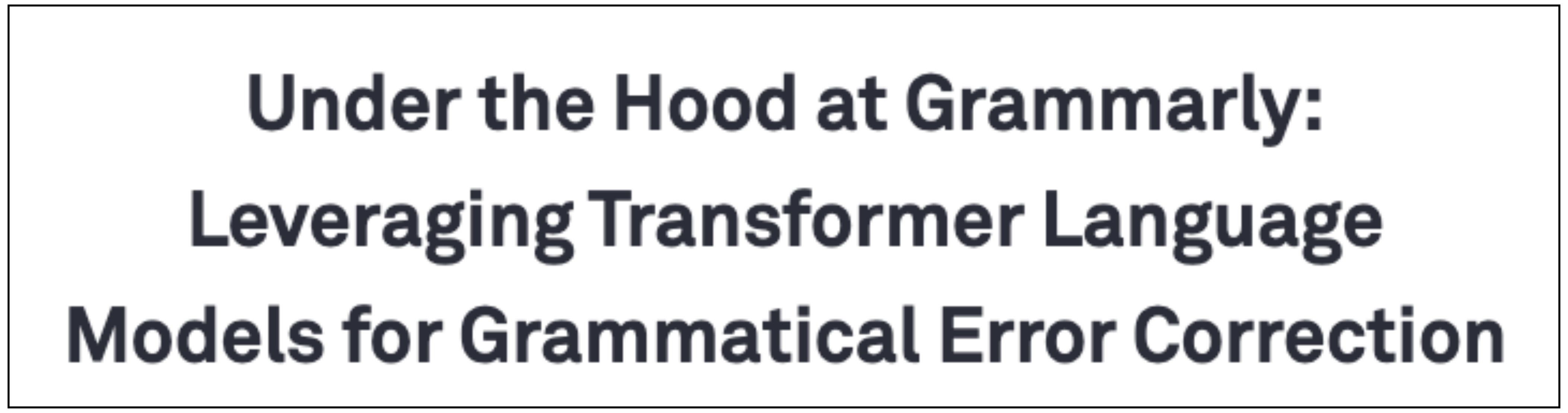
Understanding searches better than ever before - The Keyword

Oct 25, 2019 — This breakthrough was the result of **Google** research on **transformers: models** that process words in relation to all the other words in a sentence, ...

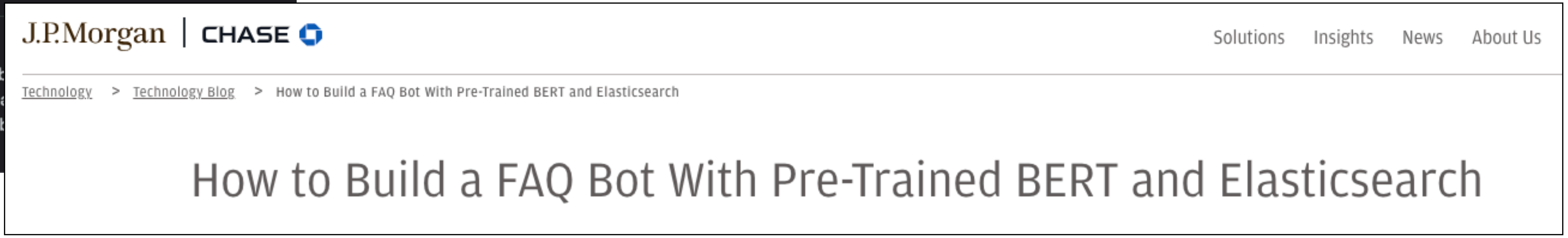
People also ask

Does Google use BERT for search?

**Google itself used BERT in its search system.** In Octob
biggest update in recent times: BERT's adoption in the sea
already adopted models to understand human language, b
one of the most significant leaps in search engine history.

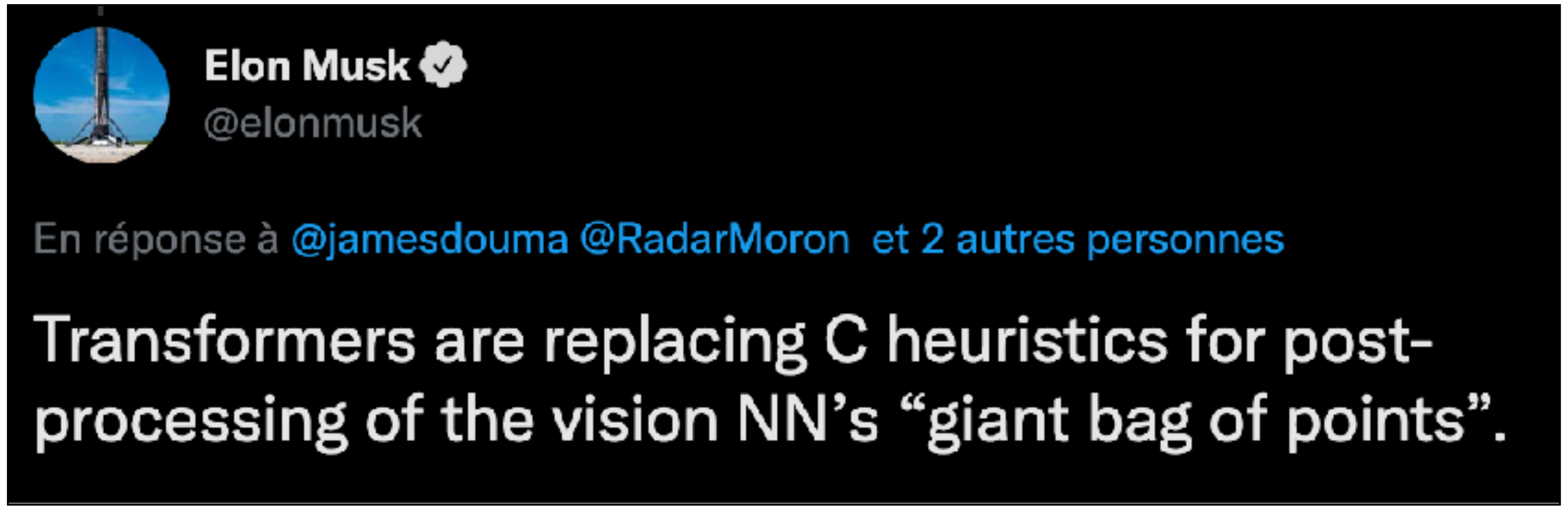


**Under the Hood at Grammarly: Leveraging Transformer Language Models for Grammatical Error Correction**



Technology > Technology Blog > How to Build a FAQ Bot With Pre-Trained BERT and Elasticsearch

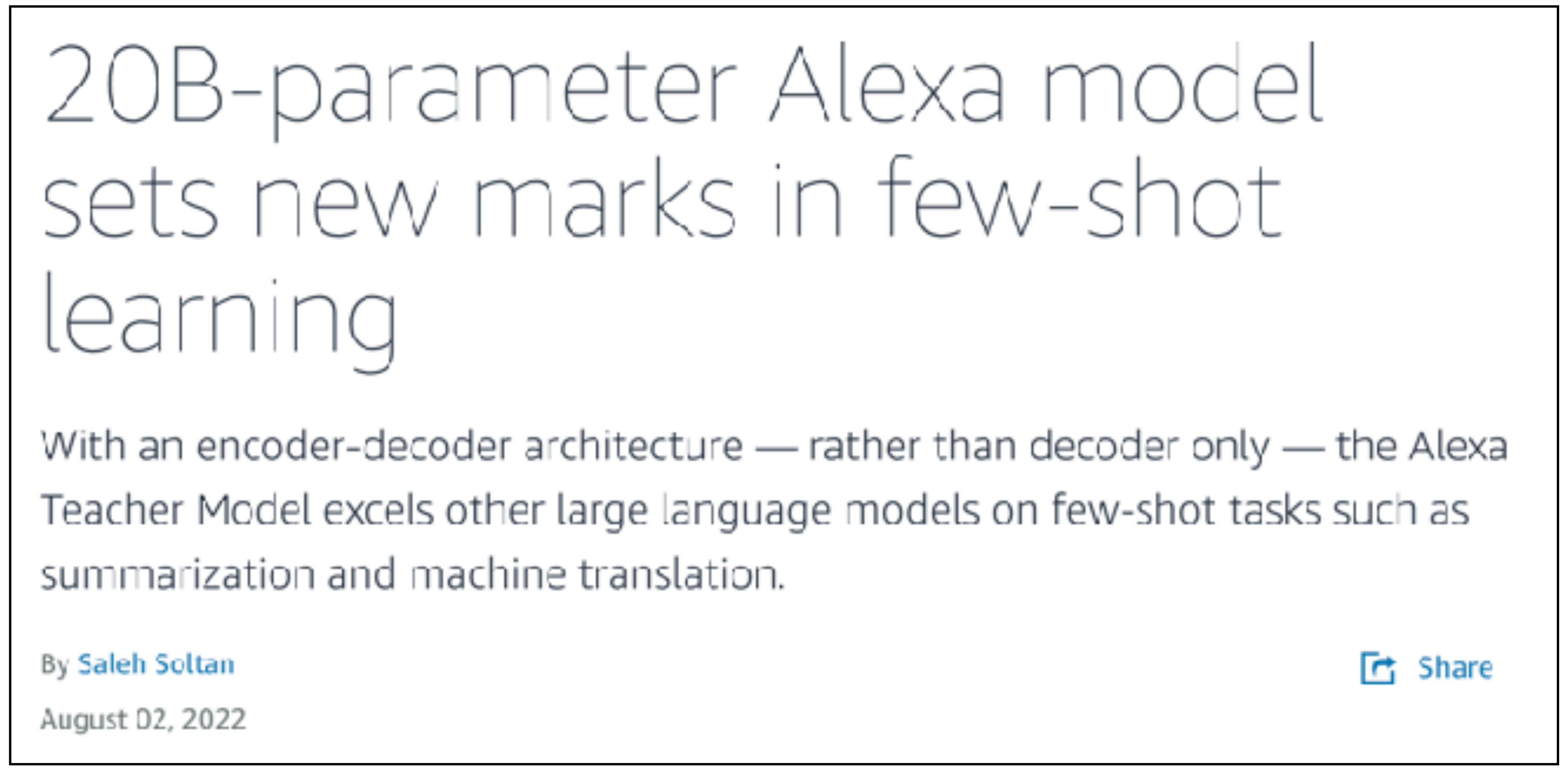
How to Build a FAQ Bot With Pre-Trained BERT and Elasticsearch



**Elon Musk**
@elonmusk

En réponse à @jamesdouma @RadarMoron et 2 autres personnes

Transformers are replacing C heuristics for post-processing of the vision NN's "giant bag of points".



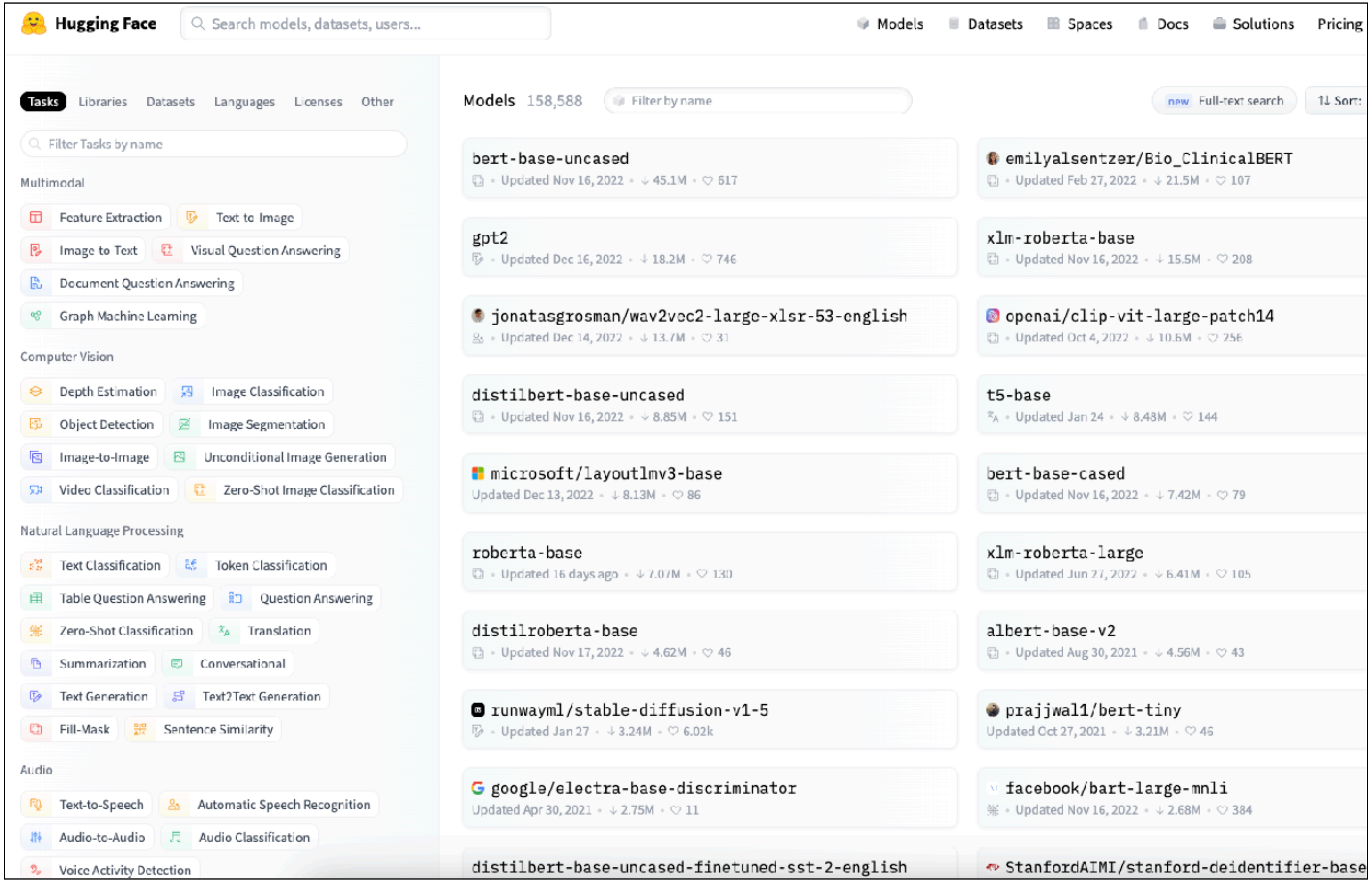
20B-parameter Alexa model sets new marks in few-shot learning

With an encoder-decoder architecture — rather than decoder only — the Alexa Teacher Model excels other large language models on few-shot tasks such as summarization and machine translation.

By Saleh Soltan
August 02, 2022

# Hugging Face: the largest collection of open source models

https://huggingface.co



**Hugging Face** Search models, datasets, users... Models Datasets Spaces Docs Solutions Pricing

Tasks Libraries Datasets Languages Licenses Other

Filter Tasks by name

Multimodal

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

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Image-to-Image
- Unconditional Image Generation
- Video Classification
- Zero-Shot Image Classification

Natural Language Processing

- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Conversational
- Text Generation
- Text2Text Generation
- Fill-Mask
- Sentence Similarity

Audio

- Text-to-Speech
- Automatic Speech Recognition
- Audio-to-Audio
- Audio Classification
- Voice Activity Detection

Models 158,588 Filter by name new Full-text search ⇅ Sort:

bert-base-uncased
Updated Nov 16, 2022 · ↓ 45.1M · ♡ 517

emilyalsentzer/Bio_ClinicalBERT
Updated Feb 27, 2022 · ↓ 21.5M · ♡ 107

gpt2
Updated Dec 16, 2022 · ↓ 18.2M · ♡ 746

xlm-roberta-base
Updated Nov 16, 2022 · ↓ 15.5M · ♡ 208

jonatasgrosman/wav2vec2-large-xlsr-53-english
Updated Dec 14, 2022 · ↓ 13.7M · ♡ 31

openai/clip-vit-large-patch14
Updated Oct 4, 2022 · ↓ 10.6M · ♡ 256

distilbert-base-uncased
Updated Nov 16, 2022 · ↓ 8.85M · ♡ 151

t5-base
Updated Jan 24 · ↓ 8.43M · ♡ 144

microsoft/layoutlmv3-base
Updated Dec 13, 2022 · ↓ 8.13M · ♡ 86

bert-base-cased
Updated Nov 16, 2022 · ↓ 7.42M · ♡ 79

roberta-base
Updated 16 days ago · ↓ 7.07M · ♡ 130

xlm-roberta-large
Updated Jun 27, 2022 · ↓ 6.41M · ♡ 105

distilroberta-base
Updated Nov 17, 2022 · ↓ 4.62M · ♡ 46

albert-base-v2
Updated Aug 30, 2021 · ↓ 4.56M · ♡ 43

runwayml/stable-diffusion-v1-5
Updated Jan 27 · ↓ 3.24M · ♡ 6.02k

prajjwal1/bert-tiny
Updated Oct 27, 2021 · ↓ 3.21M · ♡ 46

google/electra-base-discriminator
Updated Apr 30, 2021 · ↓ 2.75M · ♡ 11

facebook/bart-large-mnli
Updated Nov 16, 2022 · ↓ 2.68M · ♡ 384

distilbert-base-uncased-finetuned-sst-2-english

StanfordAIMI/stanford-deidentifier-base

220K pre-trained models (NLP, CV, Speech, etc.)

40K datasets

25+ ML libraries: Keras, spaCY, Scikit-Learn, fastai, etc.

10K organizations

500K+ users daily

# Hugging Face at a glance

Watsonx.ai (coming soon!)

Hugging Face on Azure

Amazon SageMaker

MLOps platforms

Model in production

Datasets

40,000+ datasets on the hub

Managed Inference on AWS and Azure

Inference Endpoints

Inference API

Deploy anywhere

autoTRAIN

No-code AutoML

Transformers

Diffusers

Optimum

HW-accelerated training & inference

220,000+ models on the hub

Models

Spaces

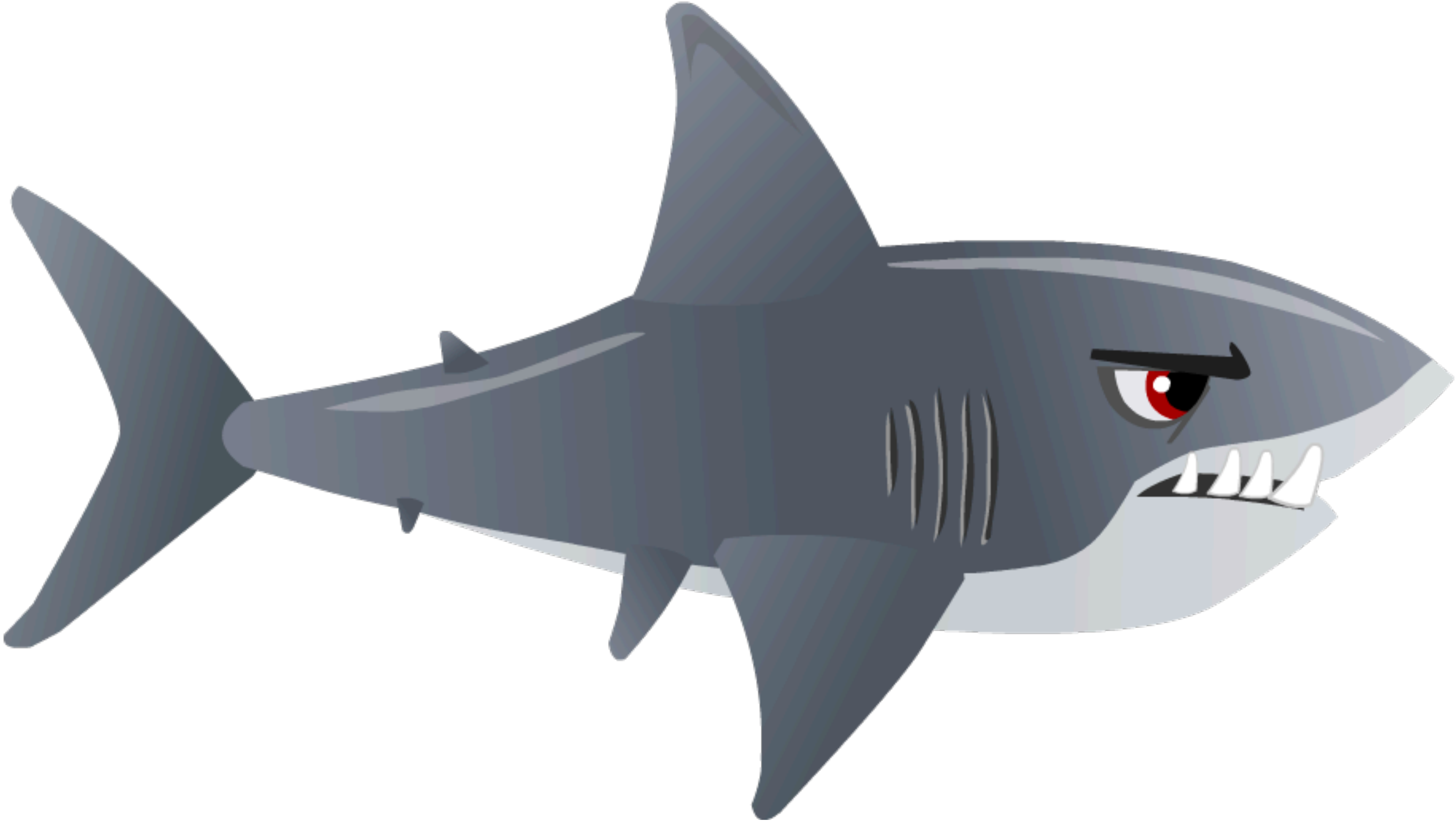
Hosted ML applications

# Expert Acceleration Program

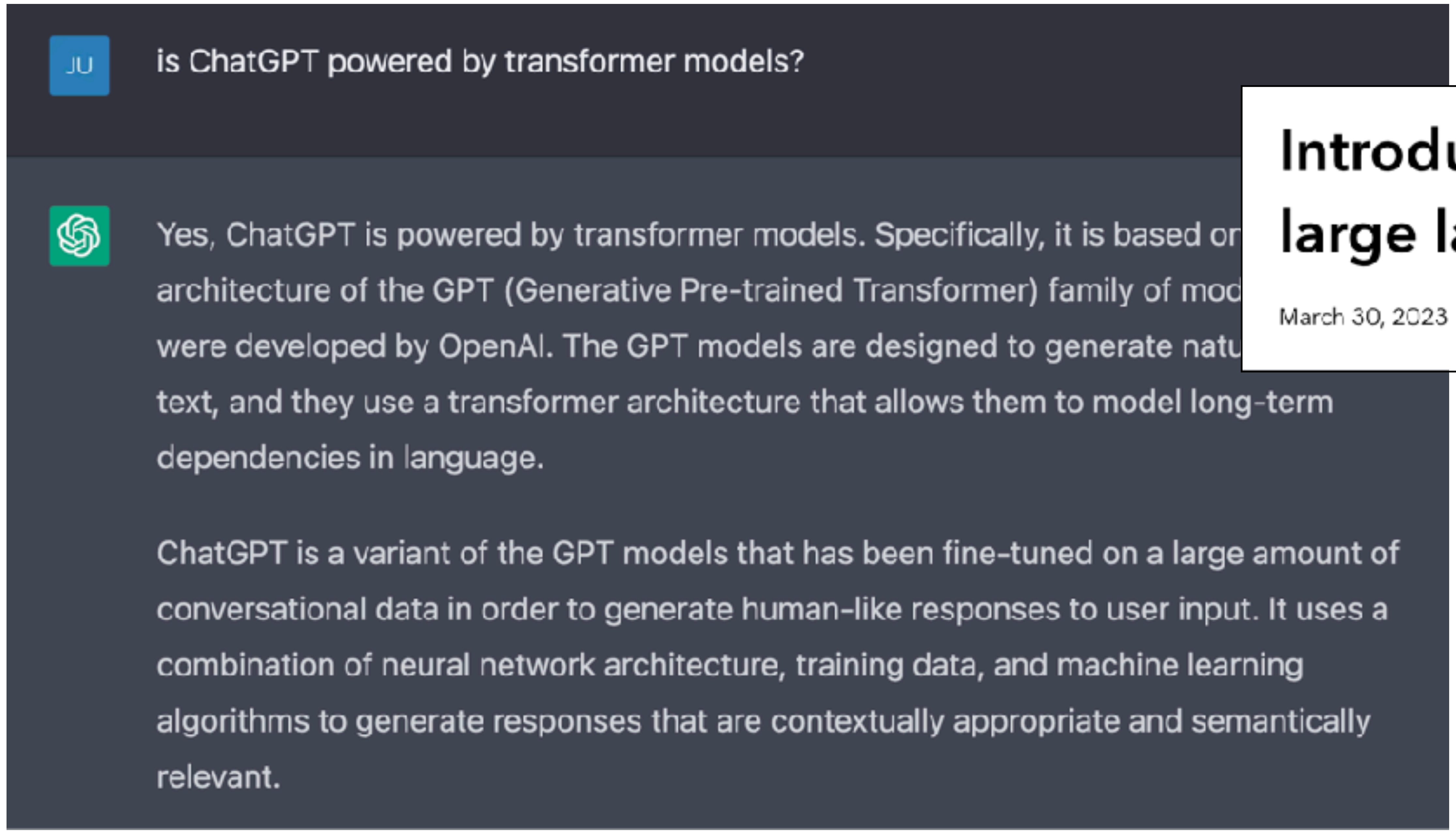
https://huggingface.co/support

- Building quality datasets for Q&A, instruction following, etc.
  - With your own data
  - With prompts/answers from other models
- Deduplicating data, a key performance factor for LLMs
- Selecting the right model architecture
- Fine-tuning models efficiently (PEFT, LoRA)
- Selecting hyper parameters
- Evaluating model quality
- Optimizing model latency in production

# 2023: Generative AI is hungry!



# Generative models in the wild



JU is ChatGPT powered by transformer models?

Yes, ChatGPT is powered by transformer models. Specifically, it is based o architecture of the GPT (Generative Pre-trained Transformer) family of mod were developed by OpenAI. The GPT models are designed to generate natu text, and they use a transformer architecture that allows them to model long-term dependencies in language.

ChatGPT is a variant of the GPT models that has been fine-tuned on a large amount of conversational data in order to generate human-like responses to user input. It uses a combination of neural network architecture, training data, and machine learning algorithms to generate responses that are contextually appropriate and semantically relevant.

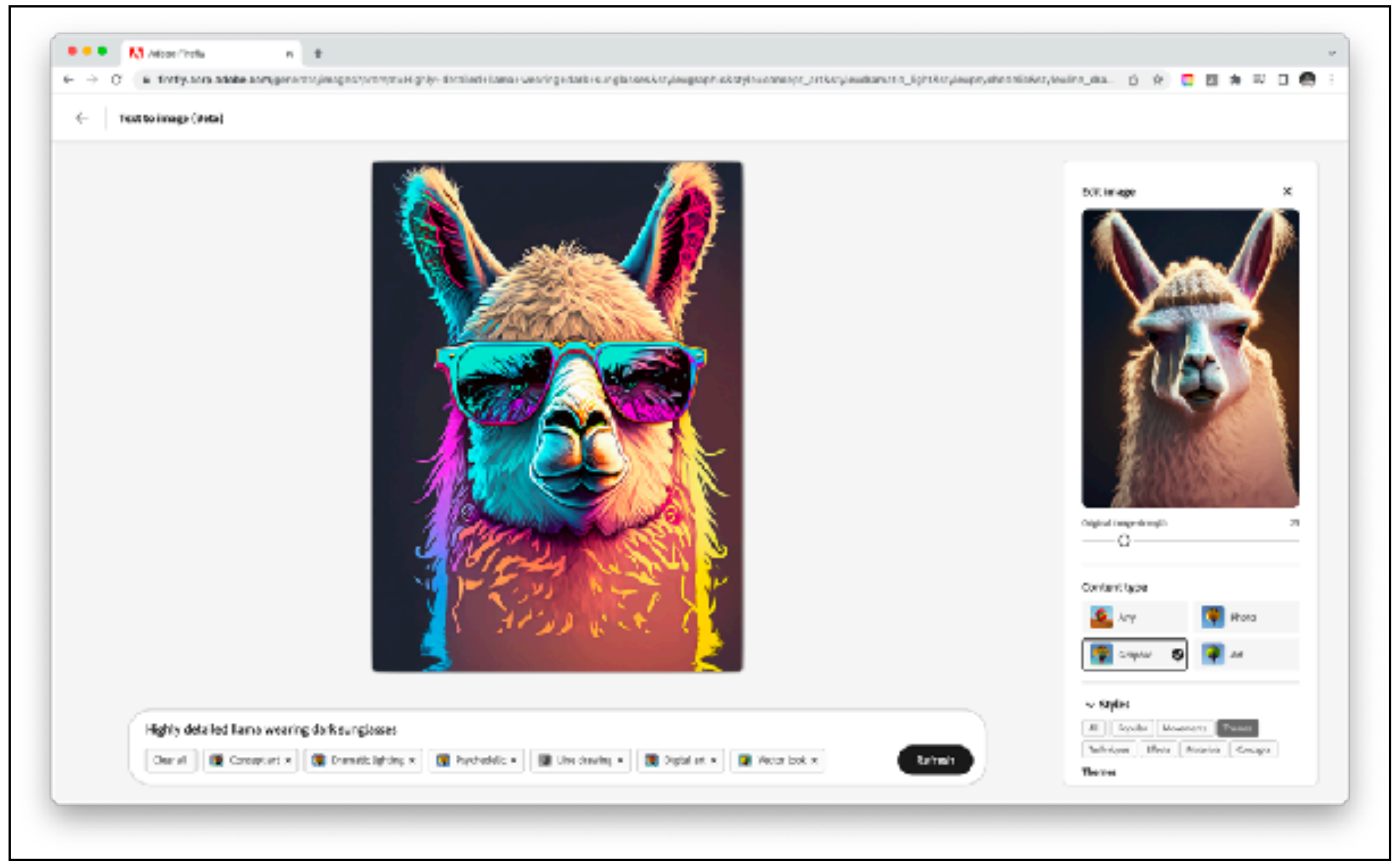
**Introducing BloombergGPT, Bloomberg's 50-billion parameter large language model, purpose-built from scratch for finance**

March 30, 2023

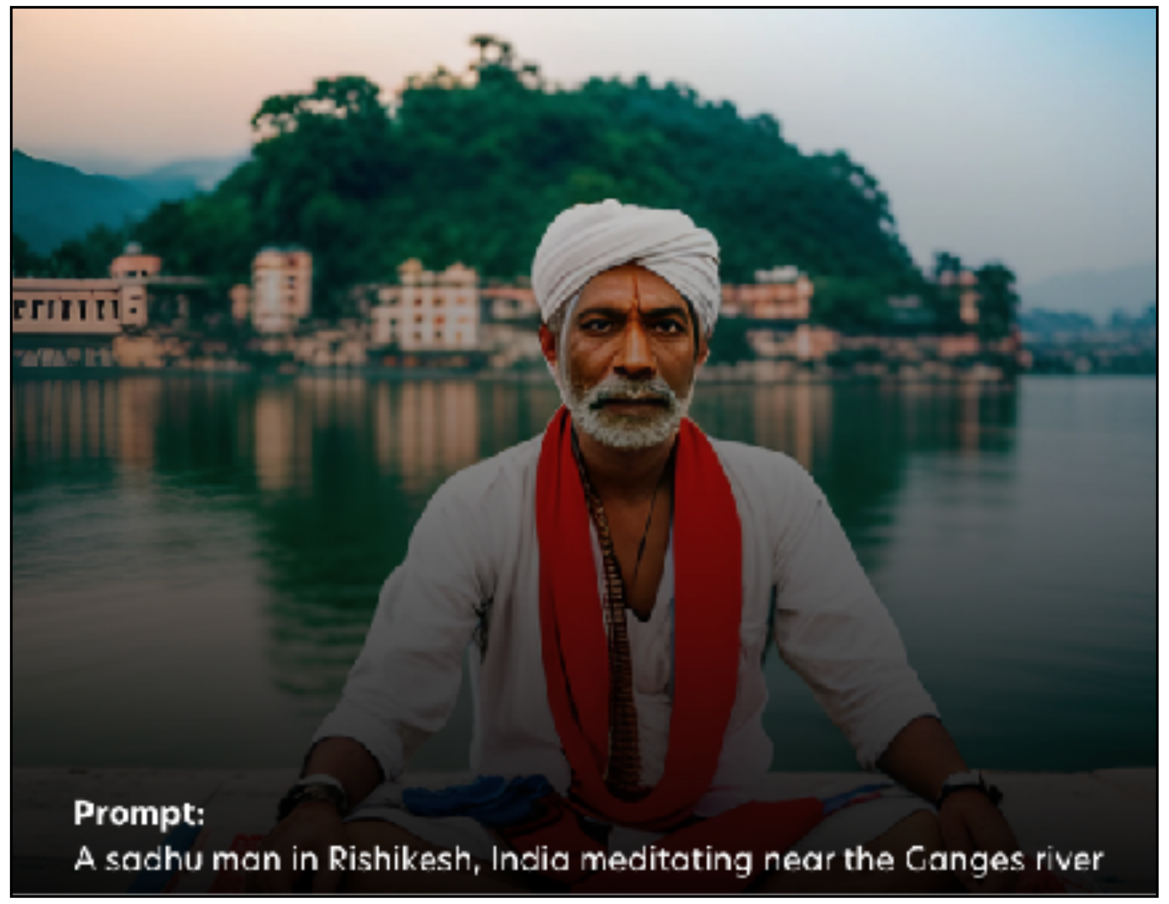
**Meet Einstein GPT, the World's First Generative AI for CRM**

**GitHub Copilot**

**Amazon CodeWhisperer**

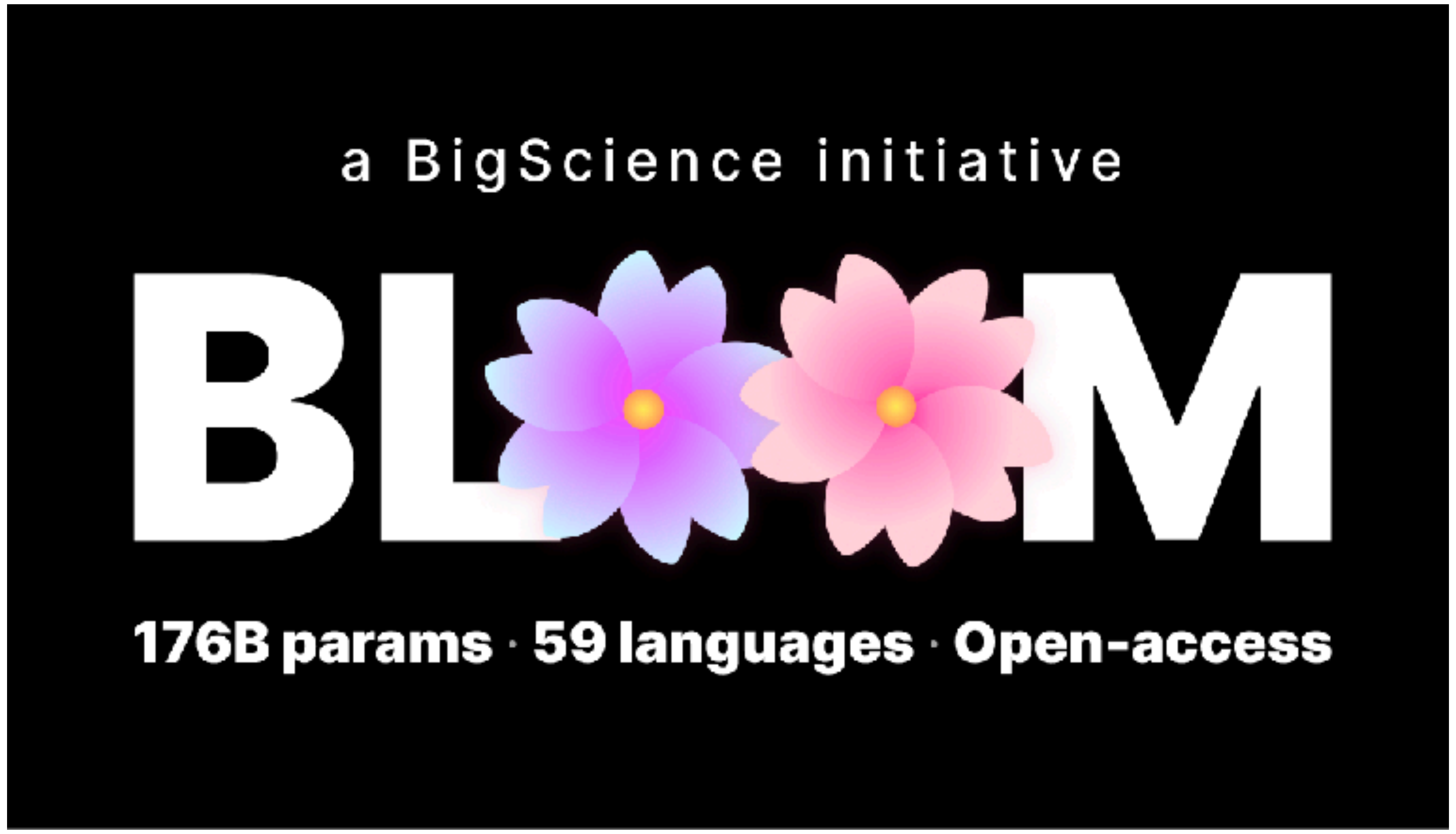


Highly detailed llama wearing dark sunglasses



**Prompt:**
A sadhu man in Rishikesh, India meditating near the Ganges river

# BLOOM: open-source alternative to GPT-3



[https://bigscience.huggingface.co](https://bigscience.huggingface.co)

[https://huggingface.co/bigscience/bloom](https://huggingface.co/bigscience/bloom)

1.5TB of text, 350B tokens

43 languages, 16 programming languages

118 days of training on 384 A100 GPUs
(public cluster)

**Smaller versions are available** : 560M, 1.1B, 1.7B, 3B, 7.1B

BLOOMZ models (same sizes) are fine-tuned for **instruction following**
[https://huggingface.co/bigscience/bloomz](https://huggingface.co/bigscience/bloomz)

# BigCode: open-source LLMs for code generation

https://www.bigcode-project.org

Dataset: https://huggingface.co/datasets/bigcode/the-stack

2.9TB of deduplicated code

Model: https://huggingface.co/bigcode/starcoder
https://arxiv.org/abs/2305.06161

15.5B parameters, 1T tokens, 80+ languages
8K context
26 days of training on 512 A100 GPUs (AWS)

# Financial LLM case study: BloombergGPT

https://arxiv.org/abs/2303.17564

- Bloomberg is a long-time customer of Hugging Face
- They built a 700B token dataset (general-purpose and financial)
- They evaluated different models and picked BLOOM as a starting point
- Based on dataset size and compute budget, they rescaled BLOOM to an optimal 50B parameters
- They used our Expert Acceleration Program (EAP) to get deep, first-hand expertise on customizing BLOOM for their own purposes: model architecture, hyper parameter selection, etc.
- They trained the model on AWS (64 instances, 512 A100 GPUs) for 52 days.

# Open Large Language Model leaderboard

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

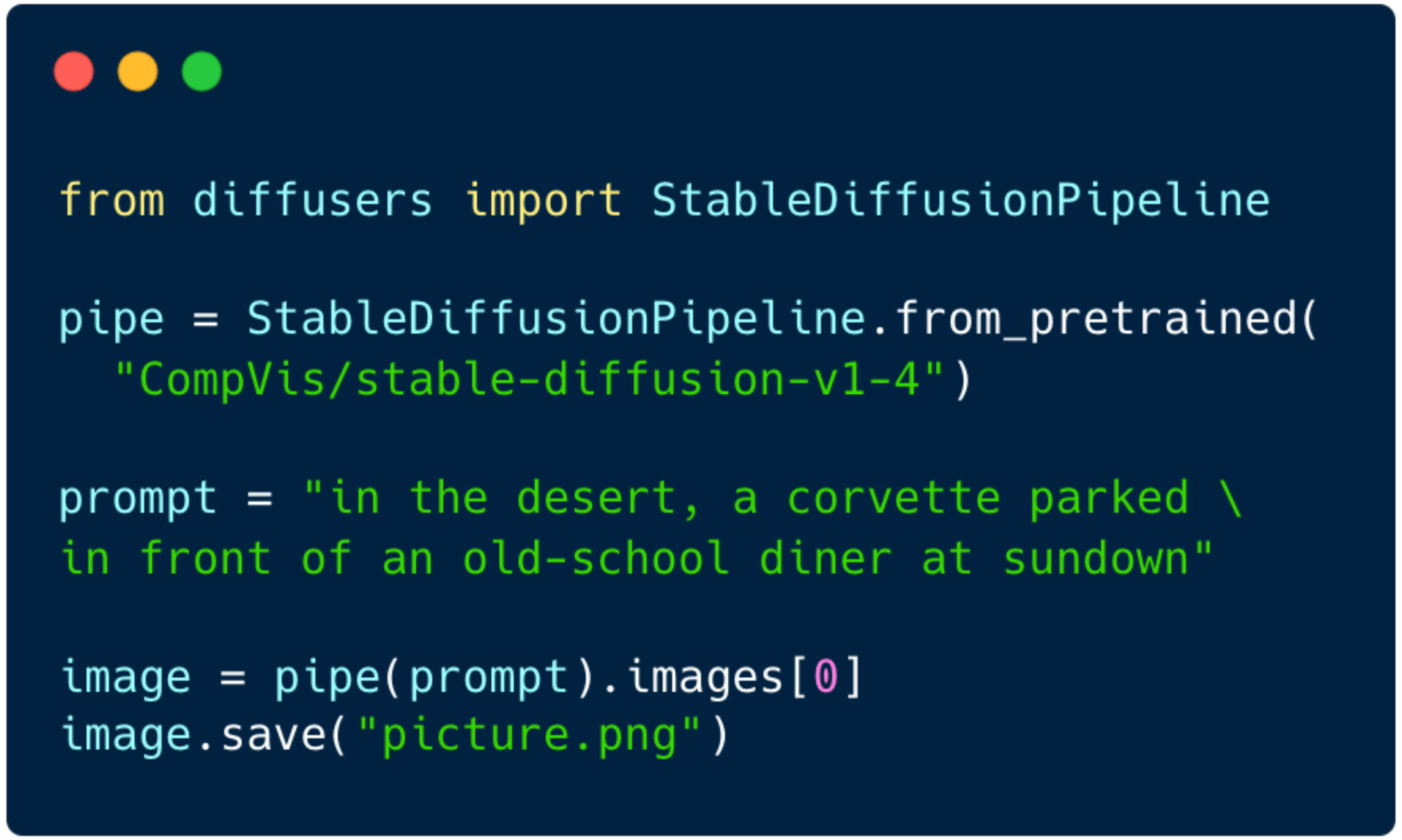
| Model ▲ | Revision ▲ | Average ⬆️ ▲ | ARC (25-shot) ⬆️ ▲ | HellaSwag (10-shot) ⬆️ ▲ | MMLU (5-shot) ⬆️ ▲ | TruthfulQA (0-shot) ⬆️ ▲ |
|---|---|---|---|---|---|---|
| tiiuae/falcon-40b-instruct | main | 63.2 | 61.6 | 84.4 | 54.1 | 52.5 |
| tiiuae/falcon-40b | main | 60.4 | 61.9 | 85.3 | 52.7 | 41.7 |
| ausboss/llama-30b-supercot | main | 59.8 | 58.5 | 82.9 | 44.3 | 53.6 |
| llama-65b | main | 58.3 | 57.8 | 84.2 | 48.8 | 42.3 |
| MetaIX/GPT4-X-Alpasta-30b | main | 57.9 | 56.7 | 81.4 | 43.6 | 49.7 |
| Aeala/VicUnlocked-alpaca-30b | main | 57.6 | 55 | 80.8 | 44 | 50.4 |
| digitous/Alpacino30b | main | 57.4 | 57.1 | 82.6 | 46.1 | 43.8 |
| Aeala/GPT4-x-AlpacaDente2-30b | main | 57.2 | 56.1 | 79.8 | 44 | 49.1 |
| TheBloke/dromedary-65b-lora-HF | main | 57 | 57.8 | 80.8 | 50.8 | 38.8 |
| TheBloke/Wizard-Vicuna-13B-Uncensored-HF | main | 57 | 53.6 | 79.6 | 42.7 | 52 |
| elinas/llama-30b-hf-transformers-4.29 | main | 56.9 | 57.1 | 82.6 | 45.7 | 42.3 |
| llama-30b | main | 56.9 | 57.1 | 82.6 | 45.7 | 42.3 |
| cyl/awsome-llama | main | 56.8 | 54.4 | 79.7 | 41.8 | 51.3 |

4 key benchmarks from the Eleuther AI Language Model Evaluation Harness

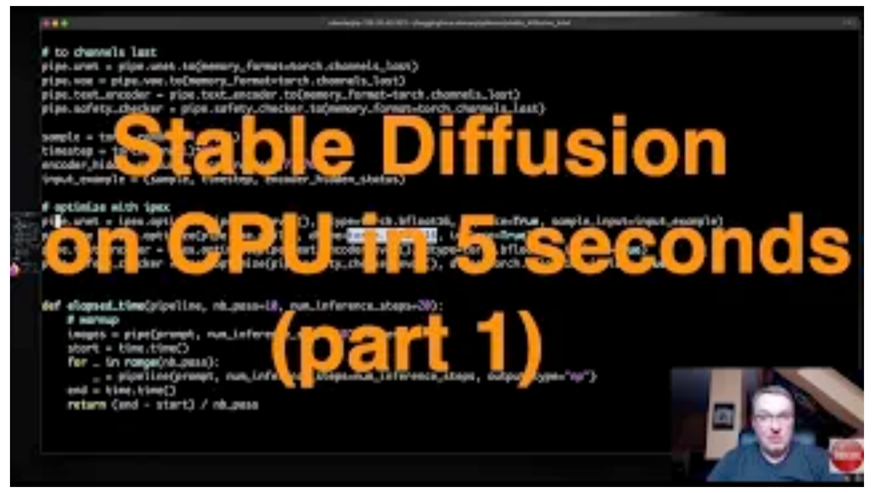
# GenAI use cases

# Example: text-to-image generation

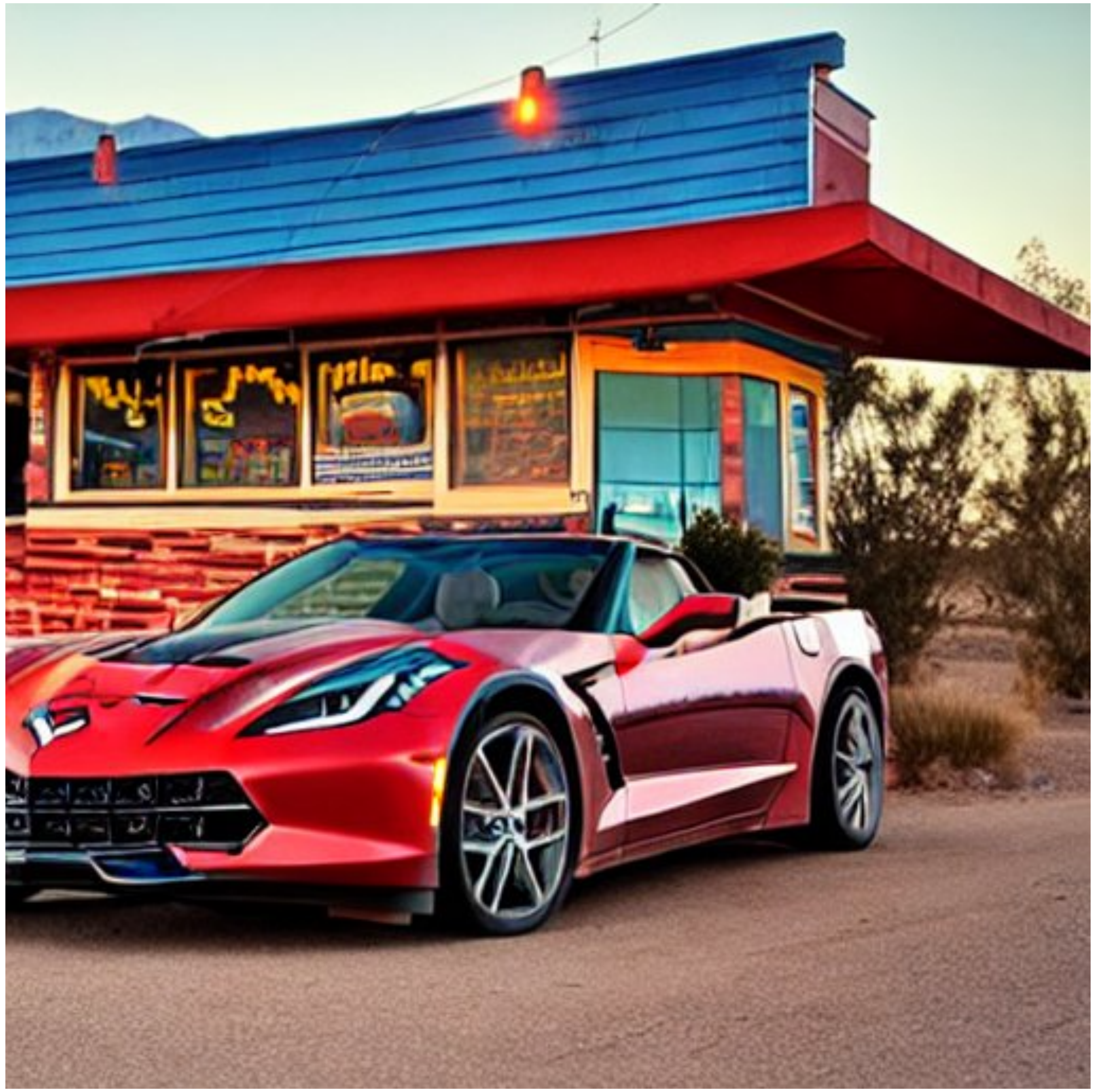
https://huggingface.co/spaces/runwayml/stable-diffusion-v1-5



```
from diffusers import StableDiffusionPipeline

pipe = StableDiffusionPipeline.from_pretrained(
  "CompVis/stable-diffusion-v1-4")

prompt = "in the desert, a corvette parked \
in front of an old-school diner at sundown"

image = pipe(prompt).images[0]
image.save("picture.png")
```

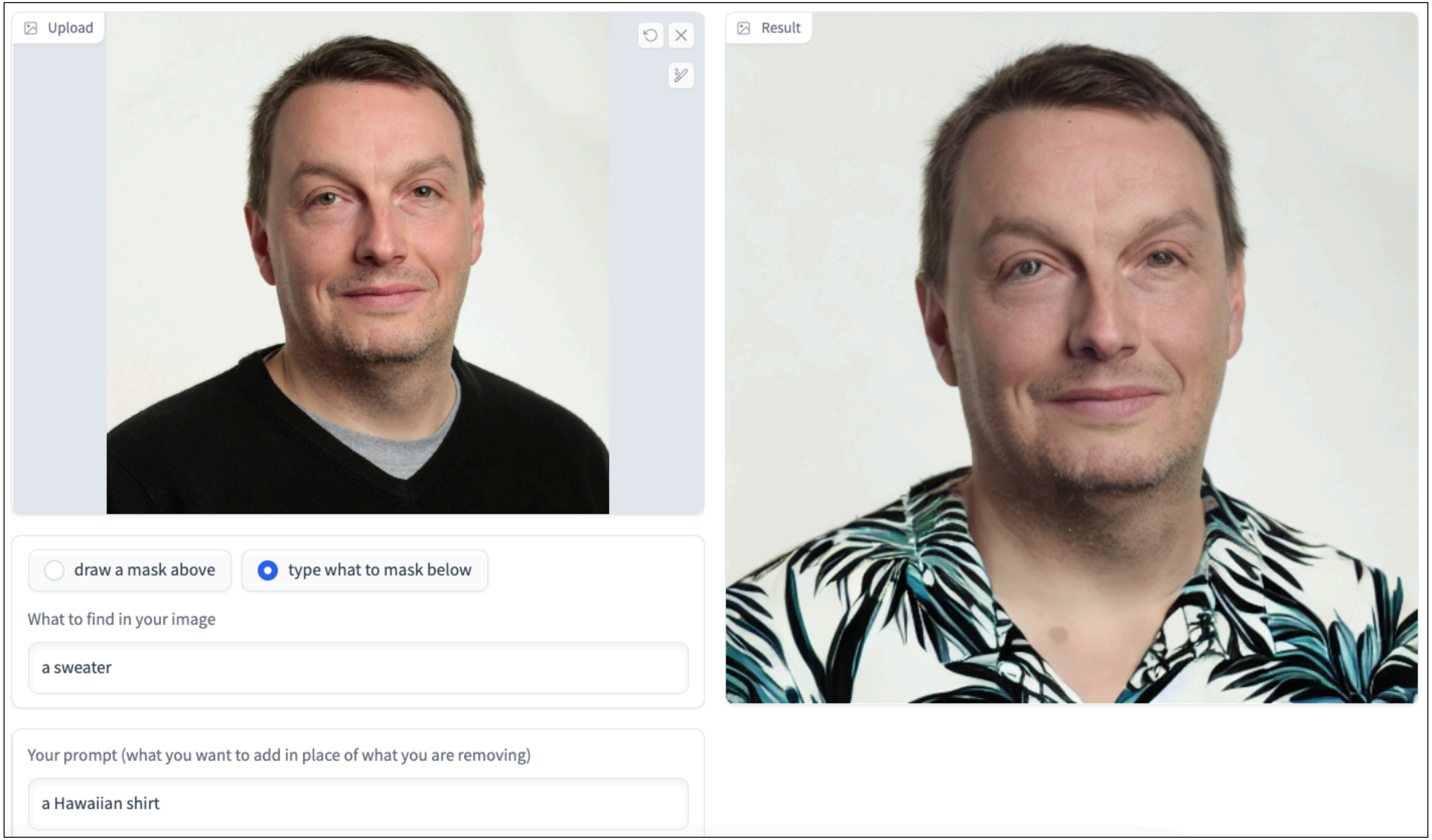


Intel Sapphire Rapids
(Amazon EC2 r7iz)
https://youtu.be/KJDCGyZ2fPw



# Example: image inpainting

https://huggingface.co/spaces/multimodalart/stable-diffusion-inpainting



○ draw a mask above ● type what to mask below

What to find in your image

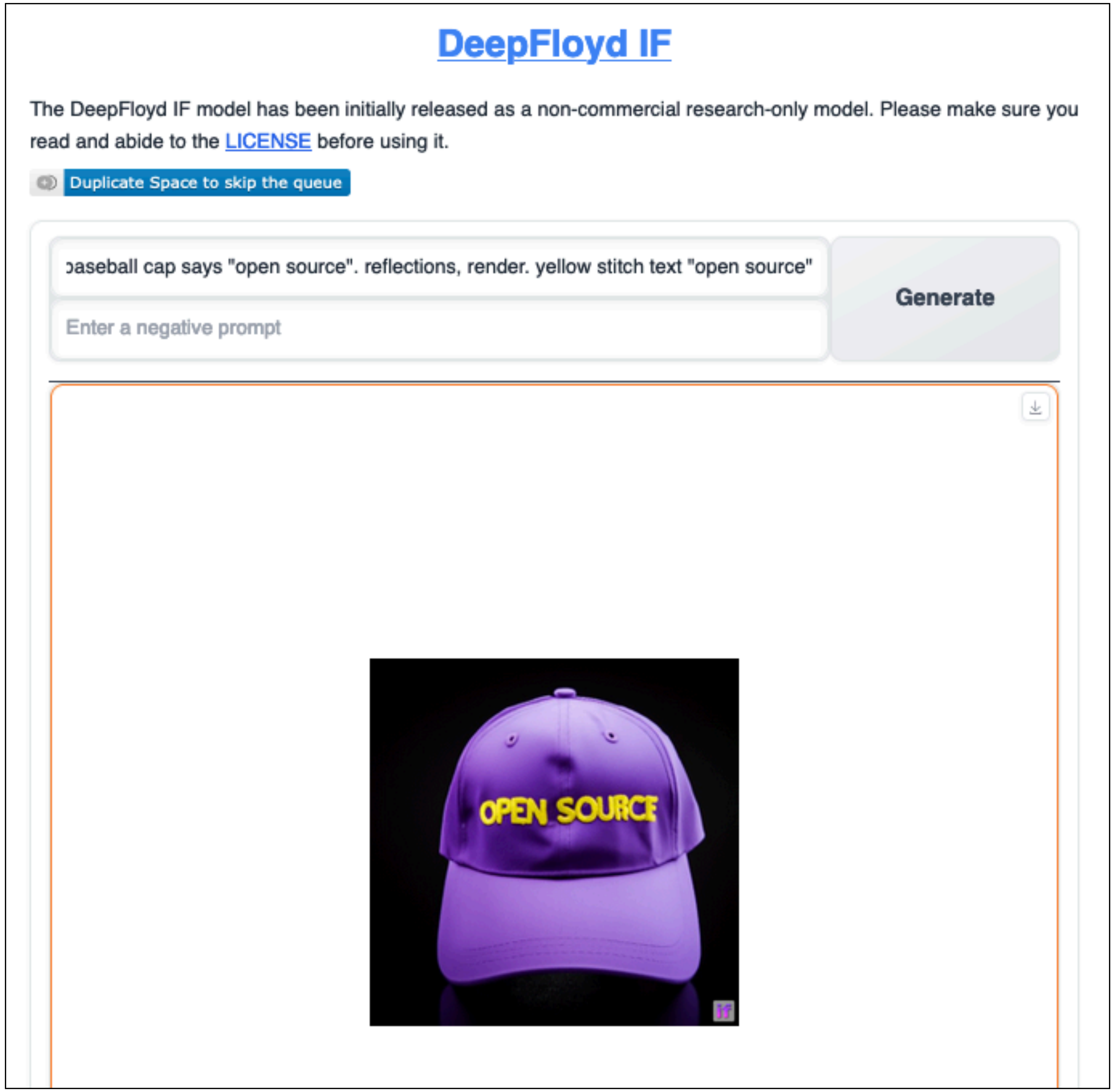
a sweater

Your prompt (what you want to add in place of what you are removing)

a Hawaiian shirt

# Example: text-to-image generation

https://huggingface.co/spaces/DeepFloyd/IF



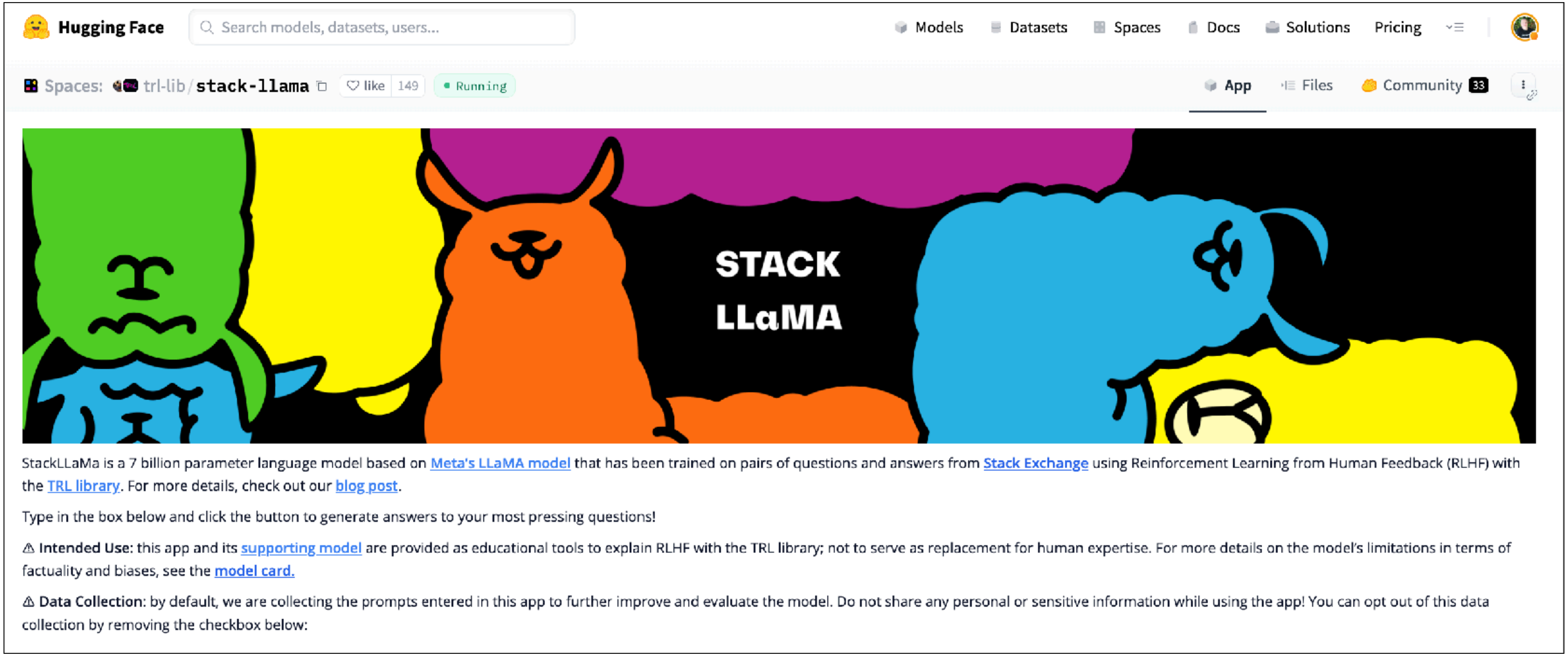
## DeepFloyd IF

The DeepFloyd IF model has been initially released as a non-commercial research-only model. Please make sure you read and abide to the LICENSE before using it.

ɔaseball cap says "open source". reflections, render. yellow stitch text "open source"

# Example: Q&A

https://huggingface.co/spaces/trl-lib/stack-llama
https://huggingface.co/blog/stackllama



StackLLaMa is a 7 billion parameter language model based on Meta's LLaMA model that has been trained on pairs of questions and answers from Stack Exchange using Reinforcement Learning from Human Feedback (RLHF) with the TRL library. For more details, check out our blog post.

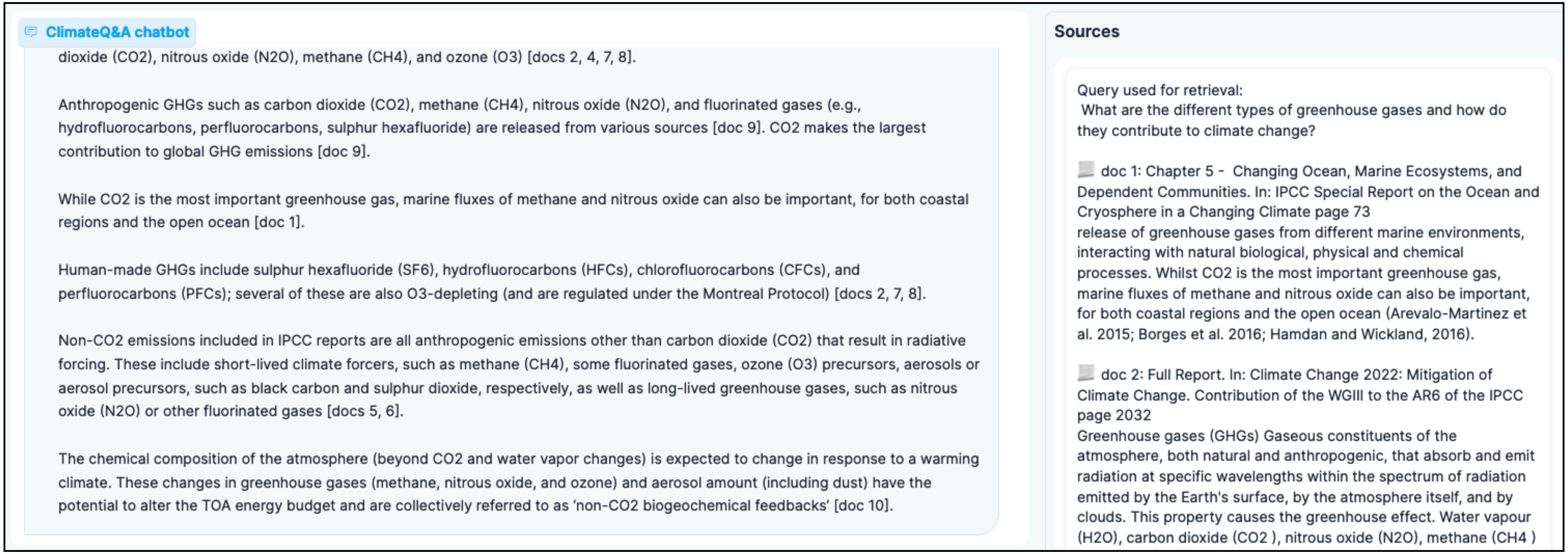
Type in the box below and click the button to generate answers to your most pressing questions!

⚠ **Intended Use:** this app and its supporting model are provided as educational tools to explain RLHF with the TRL library; not to serve as replacement for human expertise. For more details on the model's limitations in terms of factuality and biases, see the model card.

⚠ **Data Collection:** by default, we are collecting the prompts entered in this app to further improve and evaluate the model. Do not share any personal or sensitive information while using the app! You can opt out of this data collection by removing the checkbox below:

# Example: retrieval-augmented generation

https://huggingface.co/spaces/Ekimetrics/climate-question-answering



dioxide (CO2), nitrous oxide (N2O), methane (CH4), and ozone (O3) [docs 2, 4, 7, 8].

Anthropogenic GHGs such as carbon dioxide (CO2), methane (CH4), nitrous oxide (N2O), and fluorinated gases (e.g., hydrofluorocarbons, perfluorocarbons, sulphur hexafluoride) are released from various sources [doc 9]. CO2 makes the largest contribution to global GHG emissions [doc 9].

While CO2 is the most important greenhouse gas, marine fluxes of methane and nitrous oxide can also be important, for both coastal regions and the open ocean [doc 1].

Human-made GHGs include sulphur hexafluoride (SF6), hydrofluorocarbons (HFCs), chlorofluorocarbons (CFCs), and perfluorocarbons (PFCs); several of these are also O3-depleting (and are regulated under the Montreal Protocol) [docs 2, 7, 8].

Non-CO2 emissions included in IPCC reports are all anthropogenic emissions other than carbon dioxide (CO2) that result in radiative forcing. These include short-lived climate forcers, such as methane (CH4), some fluorinated gases, ozone (O3) precursors, aerosols or aerosol precursors, such as black carbon and sulphur dioxide, respectively, as well as long-lived greenhouse gases, such as nitrous oxide (N2O) or other fluorinated gases [docs 5, 6].

The chemical composition of the atmosphere (beyond CO2 and water vapor changes) is expected to change in response to a warming climate. These changes in greenhouse gases (methane, nitrous oxide, and ozone) and aerosol amount (including dust) have the potential to alter the TOA energy budget and are collectively referred to as 'non-CO2 biogeochemical feedbacks' [doc 10].

Query used for retrieval:
What are the different types of greenhouse gases and how do they contribute to climate change?

doc 1: Chapter 5 - Changing Ocean, Marine Ecosystems, and Dependent Communities. In: IPCC Special Report on the Ocean and Cryosphere in a Changing Climate page 73
release of greenhouse gases from different marine environments, interacting with natural biological, physical and chemical processes. Whilst CO2 is the most important greenhouse gas, marine fluxes of methane and nitrous oxide can also be important, for both coastal regions and the open ocean (Arevalo-Martinez et al. 2015; Borges et al. 2016; Hamdan and Wickland, 2016).

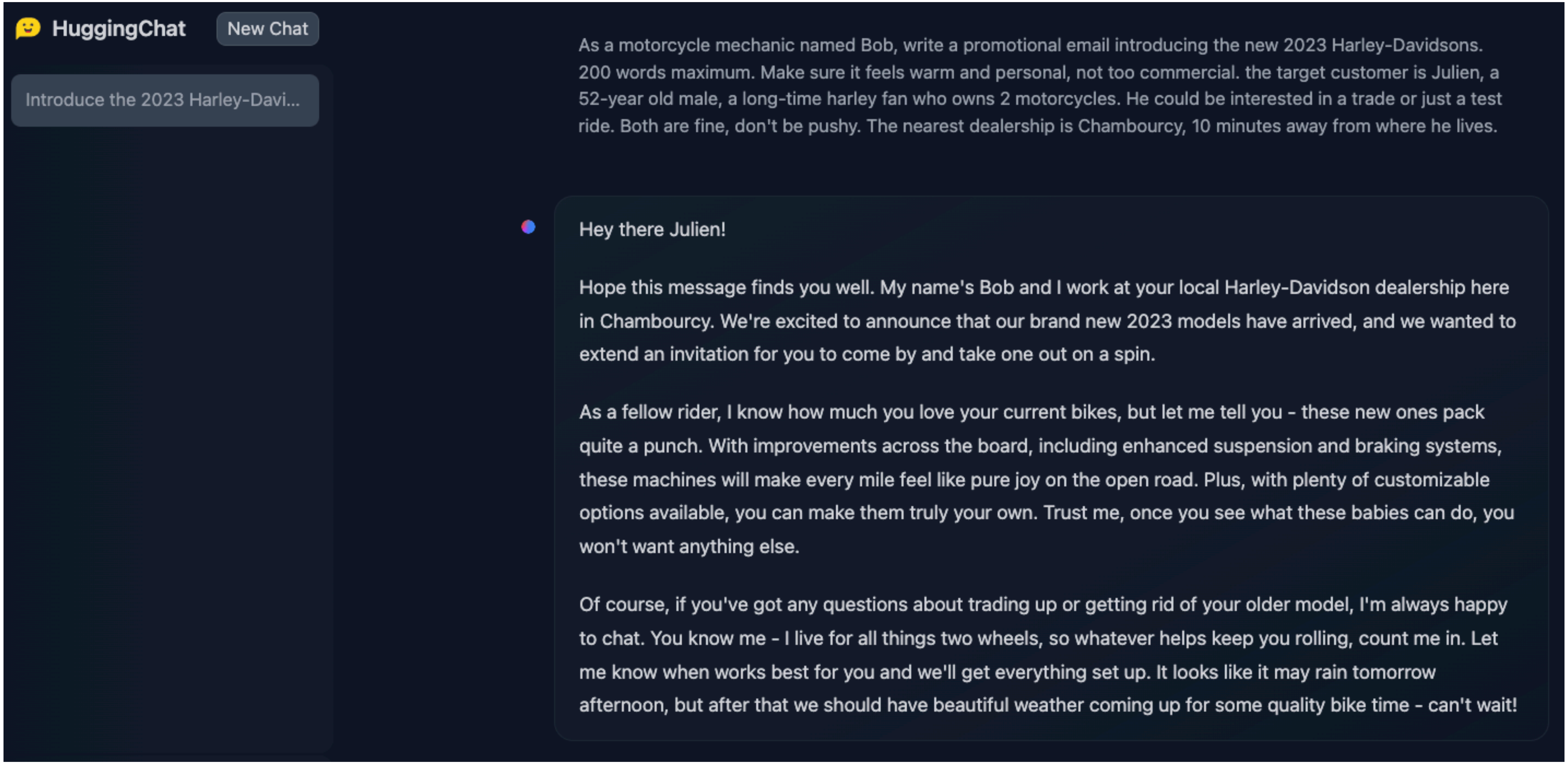
doc 2: Full Report. In: Climate Change 2022: Mitigation of Climate Change. Contribution of the WGIII to the AR6 of the IPCC page 2032
Greenhouse gases (GHGs) Gaseous constituents of the atmosphere, both natural and anthropogenic, that absorb and emit radiation at specific wavelengths within the spectrum of radiation emitted by the Earth's surface, by the atmosphere itself, and by clouds. This property causes the greenhouse effect. Water vapour (H2O), carbon dioxide (CO2 ), nitrous oxide (N2O), methane (CH4 )

Combination of <u>semantic search</u>, <u>text generation</u> and <u>prompt tweaking</u>!

This could be further improved for any domain
with a fine-tuned embedding models and/or a fine-tuned LLM

# Example: personalized marketing content with HuggingChat

https://huggingface.co/chat



As a motorcycle mechanic named Bob, write a promotional email introducing the new 2023 Harley-Davidsons. 200 words maximum. Make sure it feels warm and personal, not too commercial. the target customer is Julien, a 52-year old male, a long-time harley fan who owns 2 motorcycles. He could be interested in a trade or just a test ride. Both are fine, don't be pushy. The nearest dealership is Chambourcy, 10 minutes away from where he lives.

Hey there Julien!

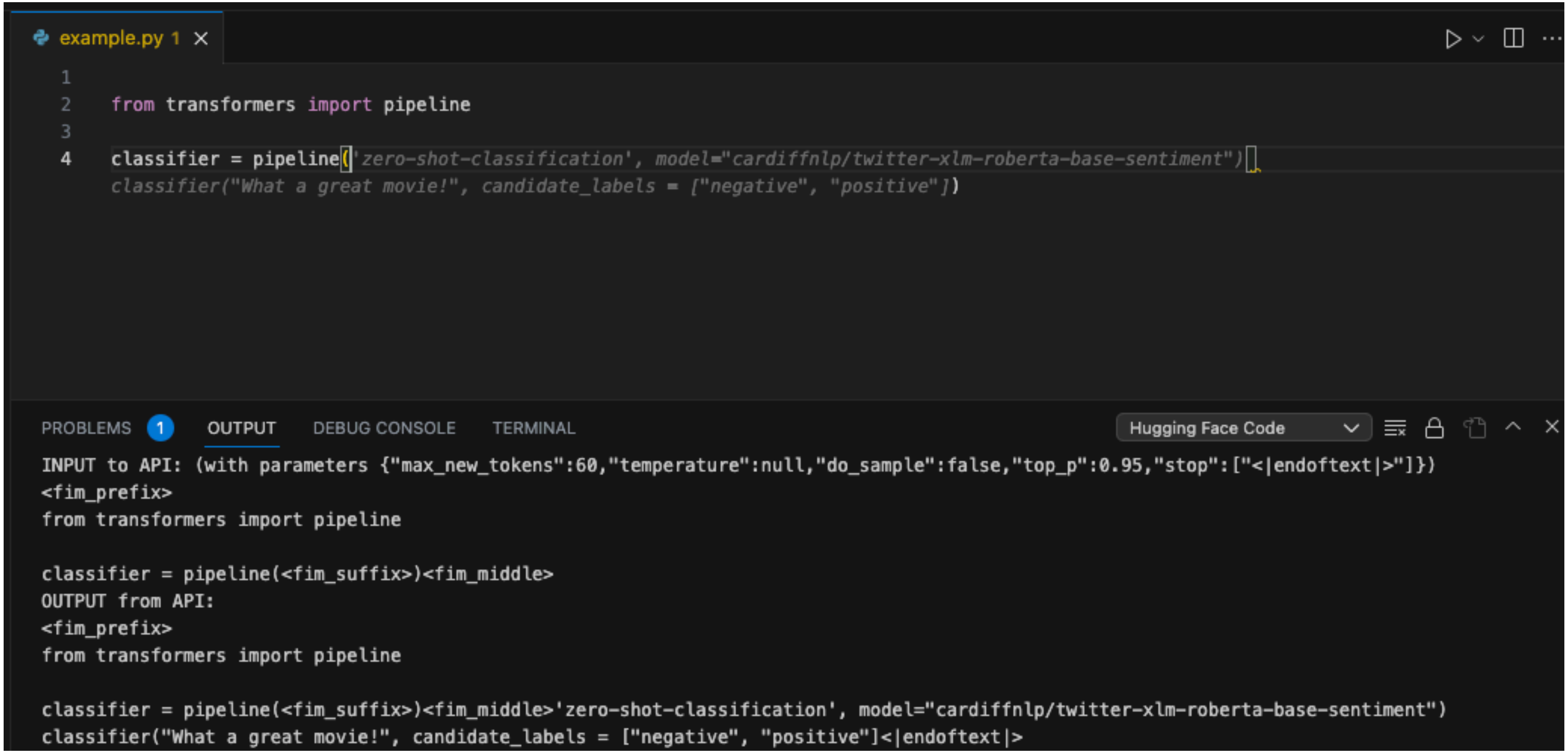
Hope this message finds you well. My name's Bob and I work at your local Harley-Davidson dealership here in Chambourcy. We're excited to announce that our brand new 2023 models have arrived, and we wanted to extend an invitation for you to come by and take one out on a spin.

As a fellow rider, I know how much you love your current bikes, but let me tell you - these new ones pack quite a punch. With improvements across the board, including enhanced suspension and braking systems, these machines will make every mile feel like pure joy on the open road. Plus, with plenty of customizable options available, you can make them truly your own. Trust me, once you see what these babies can do, you won't want anything else.

Of course, if you've got any questions about trading up or getting rid of your older model, I'm always happy to chat. You know me - I live for all things two wheels, so whatever helps keep you rolling, count me in. Let me know when works best for you and we'll get everything set up. It looks like it may rain tomorrow afternoon, but after that we should have beautiful weather coming up for some quality bike time - can't wait!

# Example: coding assistance with StarCoder

https://huggingface.co/spaces/HuggingFaceH4/starchat-playground
https://huggingface.co/blog/starchat-alpha



```python

from transformers import pipeline

classifier = pipeline('zero-shot-classification', model="cardiffnlp/twitter-xlm-roberta-base-sentiment")
classifier("What a great movie!", candidate_labels = ["negative", "positive"])
```

```
INPUT to API: (with parameters {"max_new_tokens":60,"temperature":null,"do_sample":false,"top_p":0.95,"stop":["<|endoftext|>"]})
<fim_prefix>
from transformers import pipeline

classifier = pipeline(<fim_suffix>)<fim_middle>
OUTPUT from API:
<fim_prefix>
from transformers import pipeline

classifier = pipeline(<fim_suffix>)<fim_middle>'zero-shot-classification', model="cardiffnlp/twitter-xlm-roberta-base-sentiment")
classifier("What a great movie!", candidate_labels = ["negative", "positive"]<|endoftext|>
```

# Hugging Face on AWS

# Hugging Face on AWS

Hugging Face models, datasets, and libraries

## Hugging Face Expert Acceleration Program (EAP)

awsmarketplace

**Experiment on Hugging Face**

Hugging Face Spaces

awsmarketplace Q2'23

**Train and deploy on Amazon EC2**

Hugging Face DLAMI

awsmarketplace

**Train and deploy on Amazon SageMaker**

Hugging Face DLCs
SageMaker JumpStart

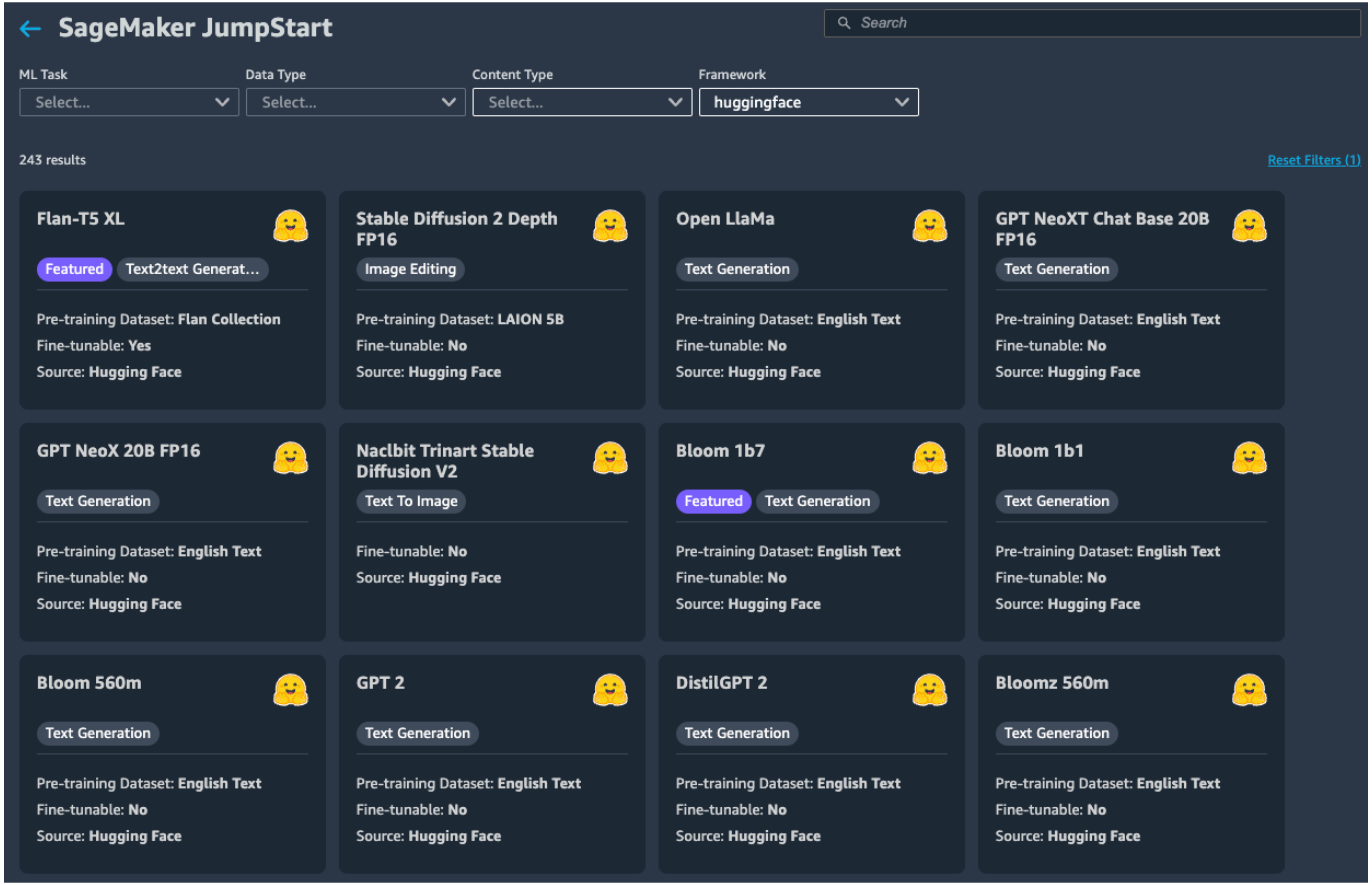
**Deploy on Hugging Face**

Hugging Face Inference Endpoints

awsmarketplace Q2'23

## AWS Infrastructure (CPU, GPU, Trainium, Inferentia)

# Demo: Amazon SageMaker JumpStart



**SageMaker JumpStart**

ML Task: Select...
Data Type: Select...
Content Type: Select...
Framework: huggingface

243 results

**Flan-T5 XL**
Featured · Text2text Generat...
Pre-training Dataset: **Flan Collection**
Fine-tunable: **Yes**
Source: **Hugging Face**

**Stable Diffusion 2 Depth FP16**
Image Editing
Pre-training Dataset: **LAION 5B**
Fine-tunable: **No**
Source: **Hugging Face**

**Open LlaMa**
Text Generation
Pre-training Dataset: **English Text**
Fine-tunable: **No**
Source: **Hugging Face**

**GPT NeoXT Chat Base 20B FP16**
Text Generation
Pre-training Dataset: **English Text**
Fine-tunable: **No**
Source: **Hugging Face**

**GPT NeoX 20B FP16**
Text Generation
Pre-training Dataset: **English Text**
Fine-tunable: **No**
Source: **Hugging Face**

**Naclbit Trinart Stable Diffusion V2**
Text To Image
Fine-tunable: **No**
Source: **Hugging Face**

**Bloom 1b7**
Featured · Text Generation
Pre-training Dataset: **English Text**
Fine-tunable: **No**
Source: **Hugging Face**

**Bloom 1b1**
Text Generation
Pre-training Dataset: **English Text**
Fine-tunable: **No**
Source: **Hugging Face**

**Bloom 560m**
Text Generation
Pre-training Dataset: **English Text**
Fine-tunable: **No**
Source: **Hugging Face**

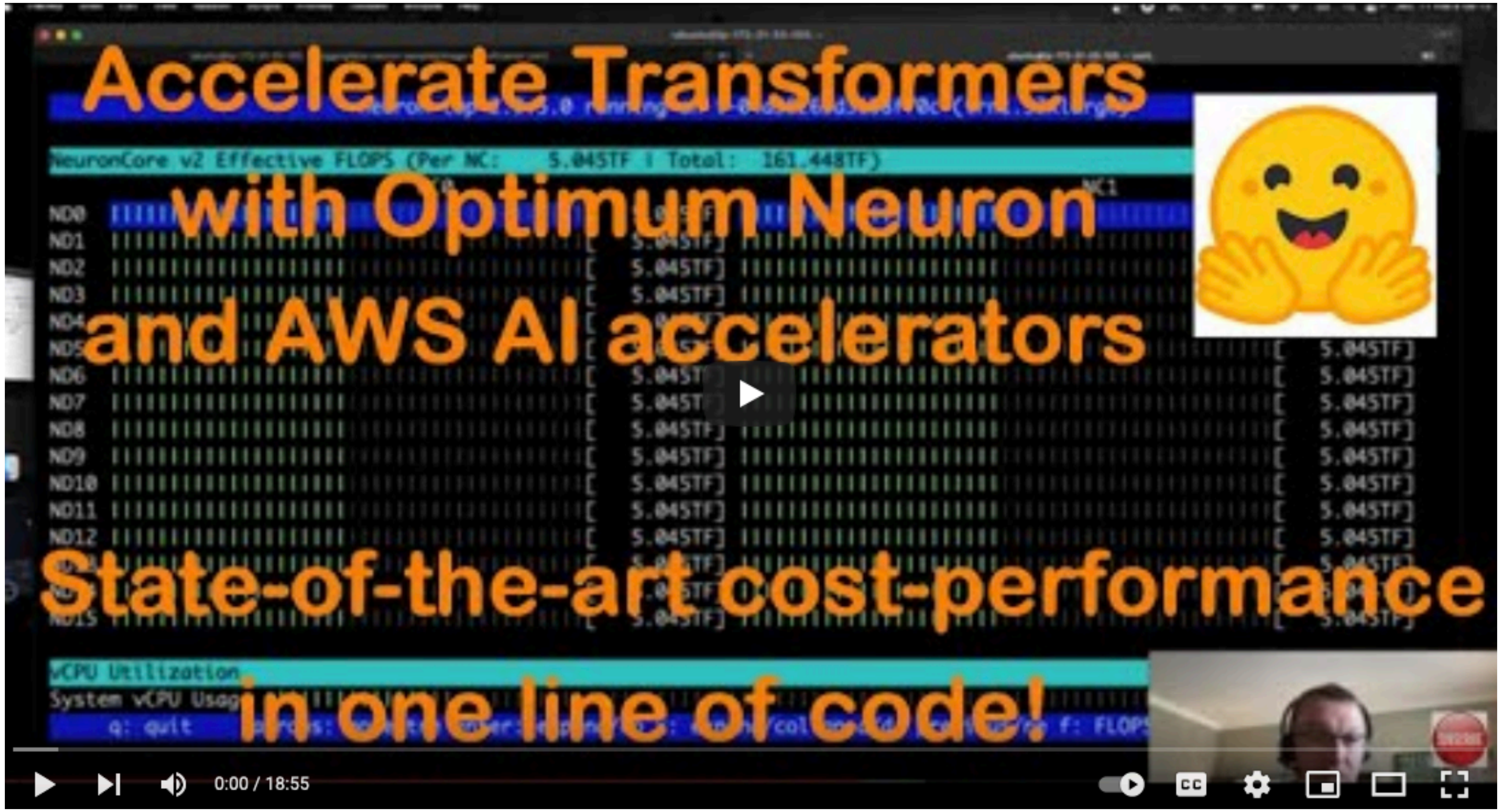
**GPT 2**
Text Generation
Pre-training Dataset: **English Text**
Fine-tunable: **No**
Source: **Hugging Face**

**DistilGPT 2**
Text Generation
Pre-training Dataset: **English Text**
Fine-tunable: **No**
Source: **Hugging Face**

**Bloomz 560m**
Text Generation
Pre-training Dataset: **English Text**
Fine-tunable: **No**
Source: **Hugging Face**

https://aws.amazon.com/blogs/machine-learning/run-text-generation-with-gpt-and-bloom-models-on-amazon-sagemaker-jumpstart

# Demo: Optimum Neuron with AWS Trainium and Inferentia2



https://www.youtube.com/watch?v=FmjTWags__Q

Vision Transformer on food101 dataset (75K training images): 1 minute/epoch
DistilBERT on 32-token sequences: 1ms P99 latency

# Conclusion

# Getting started

https://huggingface.co/tasks

https://huggingface.co/course

https://github.com/huggingface

https://huggingface.co/blog

https://huggingface.co/docs/sagemaker/index

https://www.philschmid.de/

https://youtube.com/c/juliensimonfr

# Summing things up

- AI is changing the way we build software
- Transformer models are the de facto standard for AI-powered apps.
- Don't believe the hype: today's "best" model will be superseded in weeks
- No model rules them all : find the most appropriate one for each use case
- "Small" fine-tuned open-source models are the way to go
- AWS is the best place to train and deploy transformers!