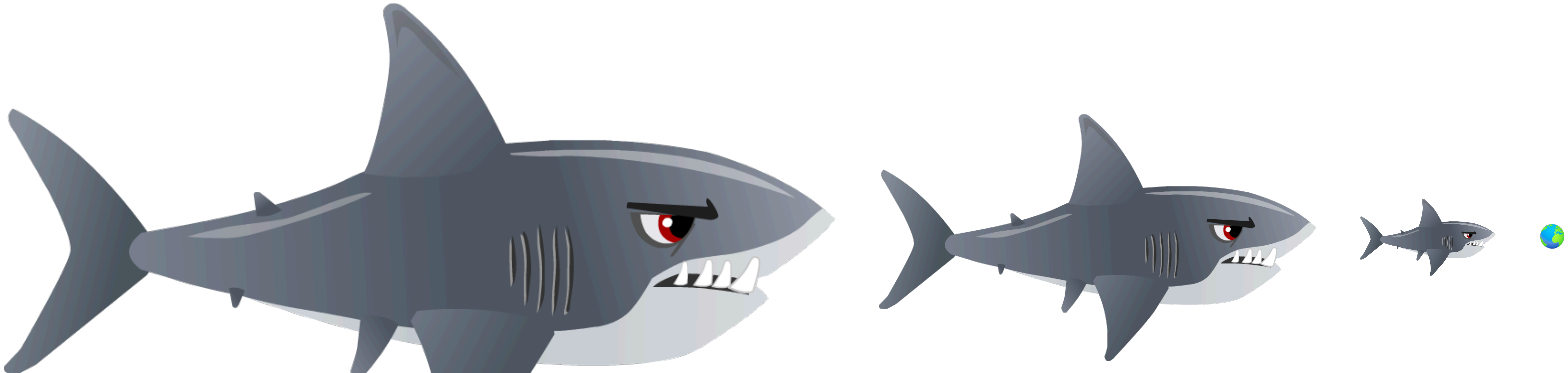
Reducing the carbon footprint of Large Language Models

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

2022: Transformers are eating Deep Learning



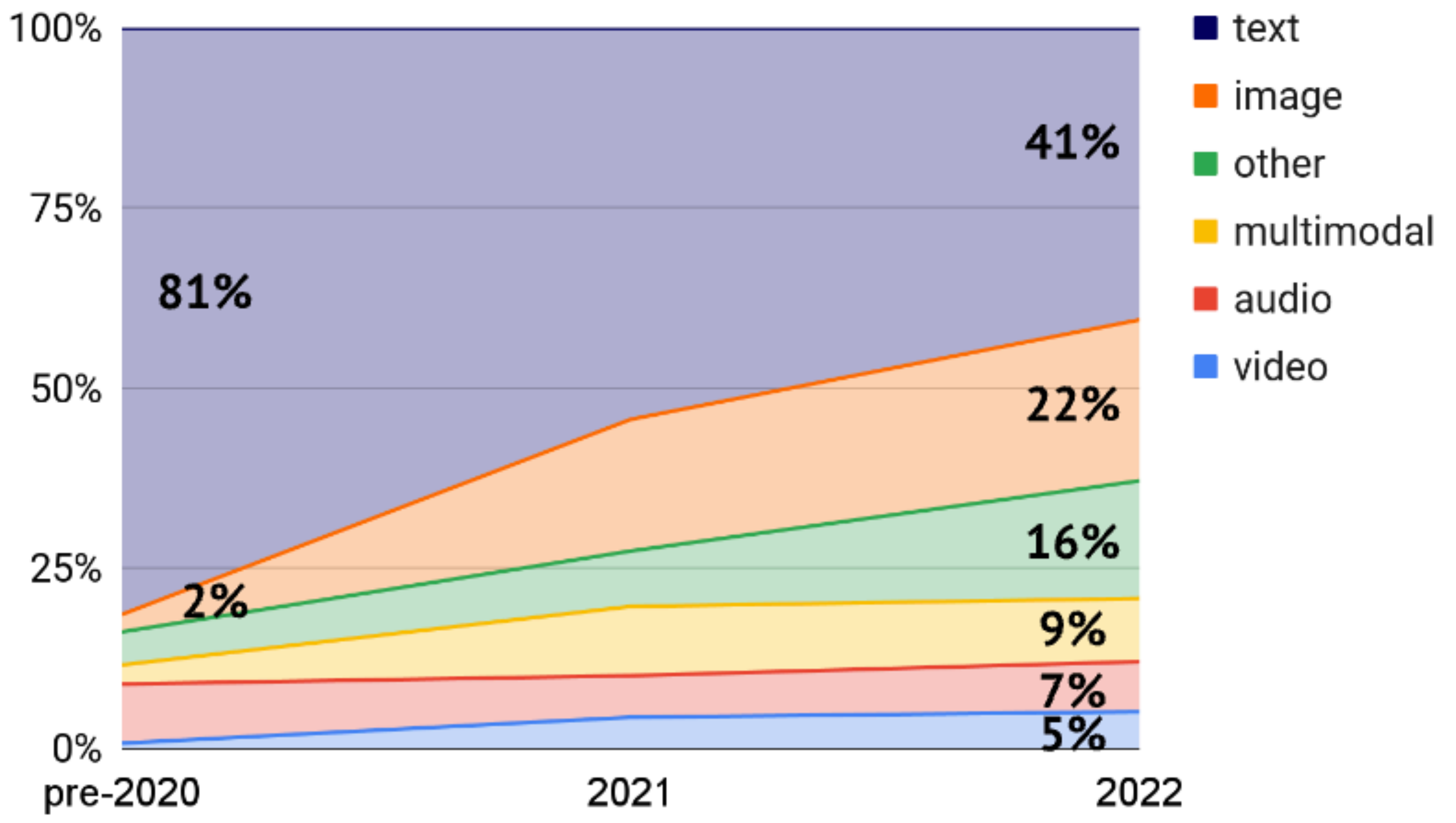
"*Transformers are emerging as a general-purpose architecture for ML*"
https://www.stateof.ai (2021)

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

All modalities, and multi-modal too

Transformers are becoming truly cross-modality

- In the 2020 State of AI Report we predicted that transformers would expand beyond NLP to achieve state of the art in computer vision. It is now clear that transformers are a candidate general purpose architecture. Analysing transformer-related papers in 2022 shows just how ubiquitous this model architecture has become.



Zeta Alpha

Transformer models in the wild

does google use transformer models for search ?

All · Images · Videos · News · Shopping · More · Tools

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 October 2019, Google announced its biggest update in recent times: BERT's adoption in the search algorithm. Google had already adopted models to understand human language, but this update was announced as one of the most significant leaps in search engine history. Nov 30, 2020

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

Share

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".

CHASE J.P.Morgan · Solutions · Insights · News · Abo

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

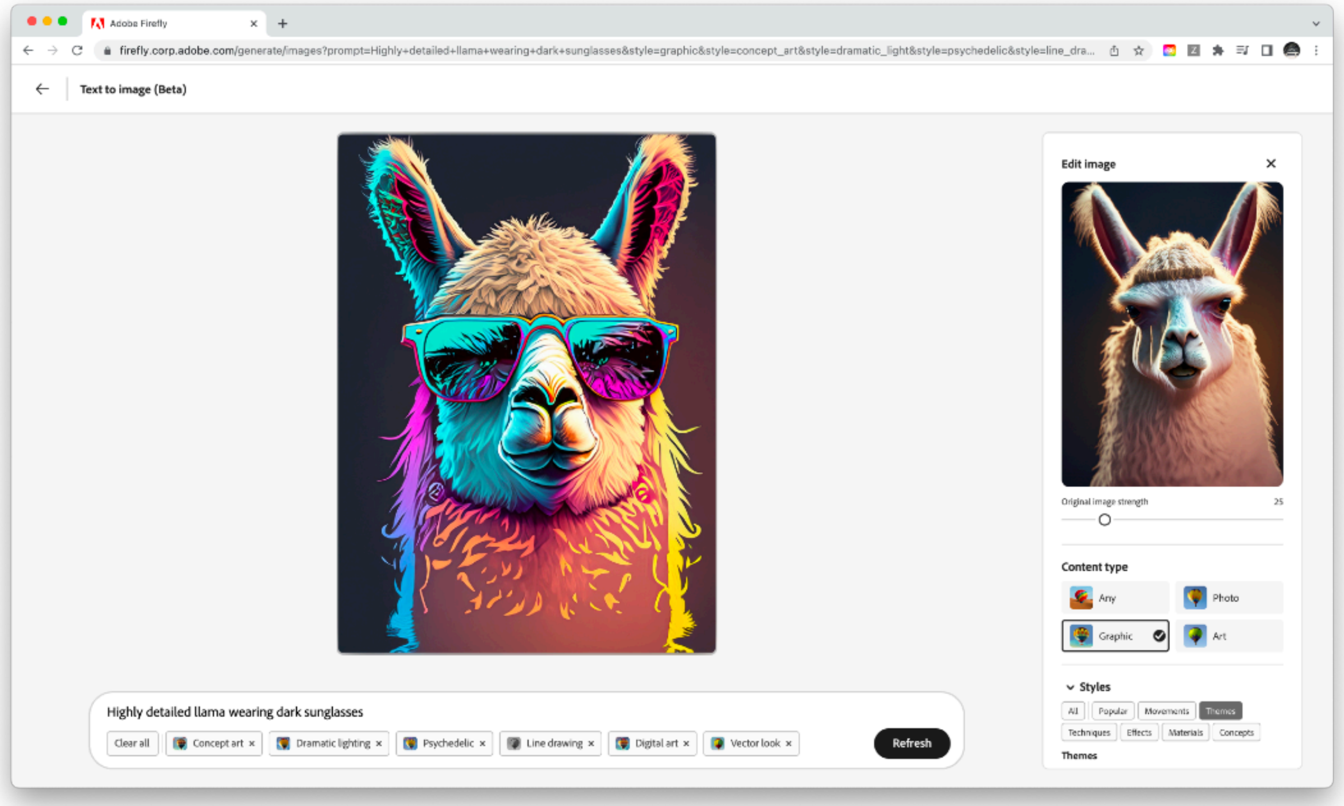
Meet Einstein GPT, the World's First Generative AI for CRM

LEARN MORE →

is ChatGPT powered by transformer models?

Yes, ChatGPT is powered by transformer models. Specifically, it is based on the architecture of the GPT (Generative Pre-trained Transformer) family of models, which were developed by OpenAI. The GPT models are designed to generate natural language 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.



2019-2020: First concerns about the impact of LLMs

Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

Abstract

Recent progress in hardware and methodology for training neural networks has ushered in a new generation of large networks trained on abundant data. These models have obtained notable gains in accuracy across many NLP tasks. However, these accuracy improvements depend on the availability of exceptionally large computational resources that necessitate similarly substantial energy consumption. As a result these models are costly to train and develop, both financially, due to the

Consumption	**CO_2e (lbs)**
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

https://arxiv.org/abs/1906.02243

The Computational Limits of Deep Learning

Neil C. Thompson[1*], Kristjan Greenewald[2], Keeheon Lee[3], Gabriel F. Manso[4]

[1]MIT Computer Science and A.I. Lab,
MIT Initiative on the Digital Economy, Cambridge, MA USA
[2]MIT-IBM Watson AI Lab, Cambridge MA, USA
[3]Underwood International College, Yonsei University, Seoul, Korea
[4] FGA, University of Brasilia, Brasilia, Brazil

*To whom correspondence should be addressed; E-mail: neil_t@mit.edu.

Abstract

Deep learning's recent history has been one of achievement: from triumphing over humans in the game of Go to world-leading performance in image classification, voice recognition, translation, and other tasks. But this progress has come with a voracious appetite for computing power. This article catalogs the extent of this dependency, showing that progress across a wide variety of applications is strongly reliant on increases in computing power. Extrapolating forward this reliance reveals that progress along current lines is rapidly becoming economically, technically, and environmentally unsustainable. Thus, continued progress in these applications will require dramatically more computationally-efficient methods, which will either have to come from changes to deep learning or from moving to other machine learning methods.

https://arxiv.org/abs/2007.05558

2022: "A Decade of Machine Learning Accelerators: Lessons Learned and Carbon Footprint" (Google)

[Str19] estimated emissions of this Neural Architecture Search (NAS)

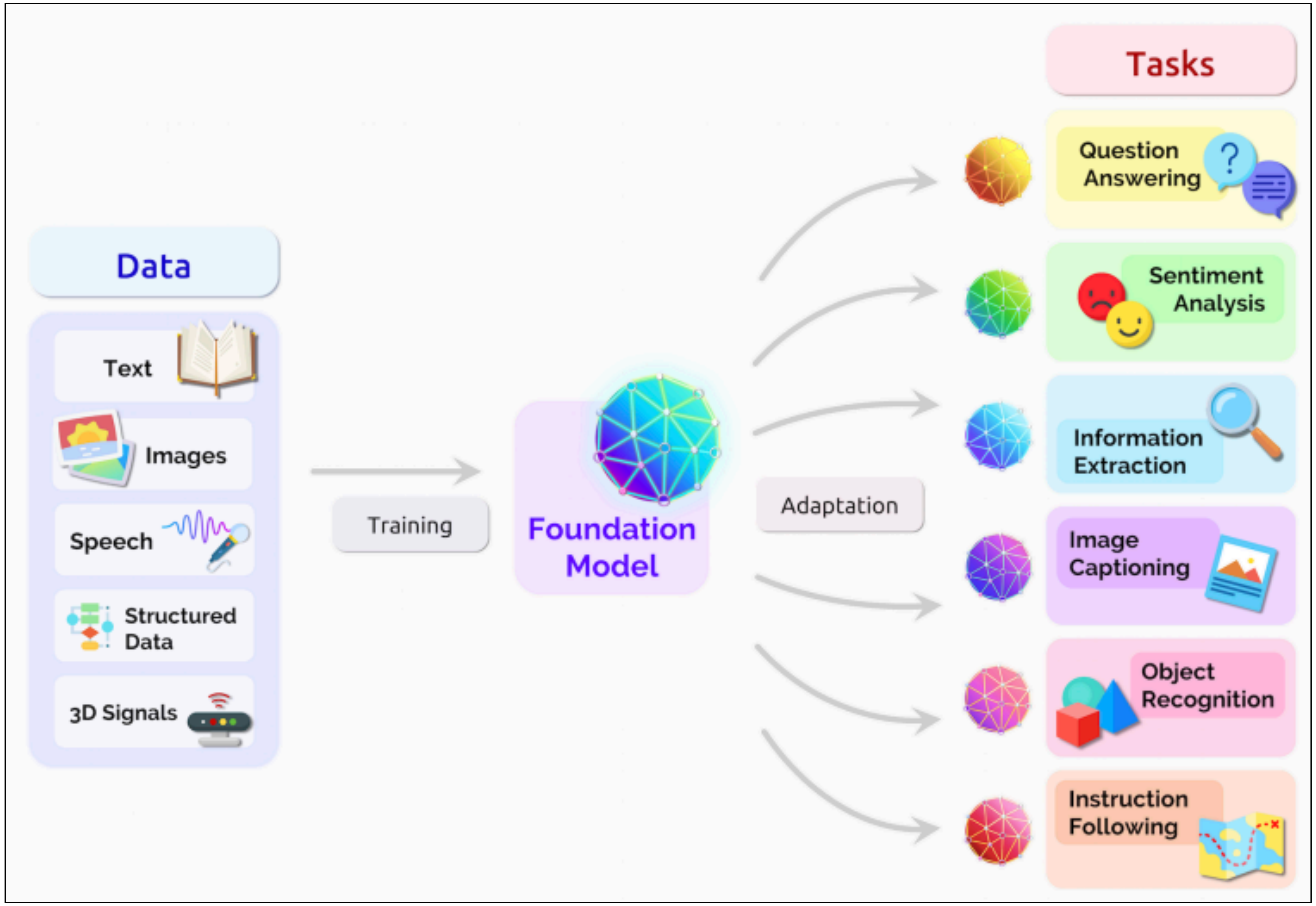
- Cited ~1500 times
- Used P100 vs TPUv2, US averages vs Google DC: 5X too high for NAS
- + Used full model vs small proxy for search: 19X ⇒ 88X too high for NAS

Some papers citing [Str19] confused NAS with Training cost

- NAS done once per problem domain+architectural search space
- NAS emissions ~1000x training emissions of DNN model found in search

https://chips-compilers-mlsys-22.github.io/

From Large Language Models to Foundation Models



Source: "On the Opportunities and Risks of Foundation Models", Stanford University

Very large models (> 10B parameters)

Unsupervised or self-supervised learning

Often trained on multimodal data

Not intended to be used directly for any particular goal

Intended to serve as a basis for downstream models specialized for particular tasks

Examples: GPT-3 (Open AI), Florence (Microsoft), Flamingo (DeepMind), LLaMA (Meta), PaLM (Google), BLOOM (Hugging Face)

Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models (2020)

D. Estimating the Energy and Carbon Footprint of GPT-3

Brown et al. (2020) report that the GPT-3 model with 175 billion parameters used $3.14 \cdot 10^{23}$ floating point operations (FPOs) of compute to train using NVIDIA V100 GPUs on a cluster provided by Microsoft. We assume that these are the most powerful V100 GPUs, the V100S PCIe model, with a tensor performance of 130 TFLOPS[12] and that the Microsoft data center has a PUE of 1.125, the average for new Microsoft data centers in 2015[13]. The compute time on a single GPU is therefore

$$\frac{3.14 \cdot 10^{23}\ \text{FPOs}}{130 \cdot 10^{12}\ \text{FLOPS}} = 2415384615.38\text{s} = 27955.84\text{d}.$$

This is equivalent to about 310 GPUs running non-stop for 90 days. If we use the thermal design power (TDP) of the V100s and the PUE, we can estimate that this used

$$\begin{aligned} 250\text{W} \cdot 2415384615.38\text{s} \cdot 1.125 &= 679326923075.63\text{J} \\ &= 188701.92\text{kWh}. \end{aligned}$$

Figure 8. Average carbon intensity ($\text{gCO}_2\text{eq/kWh}$) of EU-28 countries in 2016. The intensity is calculated as the ratio of emissions from public electricity production and gross electricity production. Data is provided by the European Environment Agency (EEA). See https://www.eea.europa.eu/ds_resolveuid/3f6dc9e9e92b45b9b829152c4e0e7ade.

Using the average carbon intensity of USA in 2017 of 449.06 $\text{gCO}_2\text{eq/kWh}$[14], we see this may emit up to

$$\begin{aligned} 449.06\text{gCO}_2\text{eq/kWh} \cdot 188701.92\text{kWh} \\ = 84738484.20\text{gCO}_2\text{eq} \\ = 84738.48\text{kgCO}_2\text{eq}. \end{aligned}$$

This is equivalent to

$$\frac{84738484.20\text{gCO}_2\text{eq}}{120.4\text{gCO}_2\text{eqkm}^{-1}} = 703808.01\text{km}$$

travelled by car using the average CO_2eq emissions of a newly registered car in the European Union in 2018[15].

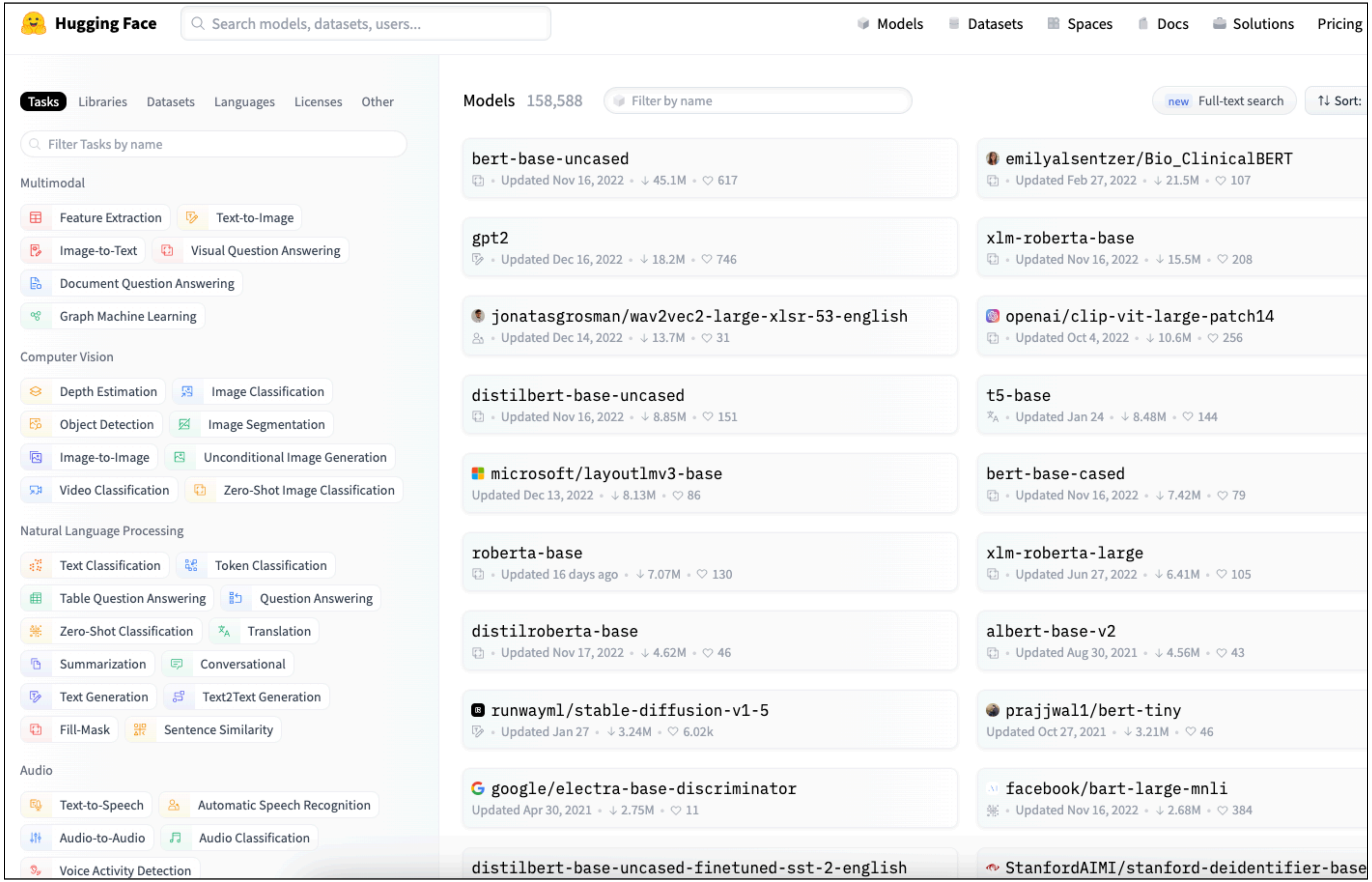
https://arxiv.org/abs/2007.03051

A single GPT-3 training run takes about a month on 1000 V100 GPUs.

In an optimized US data center, CO2 emissions are "equivalent" to driving to the Moon and back (85 tons CO2eq)

GPT-3 training on 1024 A100s is estimated at 34 days (95.4 A100-years)
https://arxiv.org/abs/2104.04473

Hugging Face: the largest collection of open source models



https://huggingface.co

170K models

28K datasets

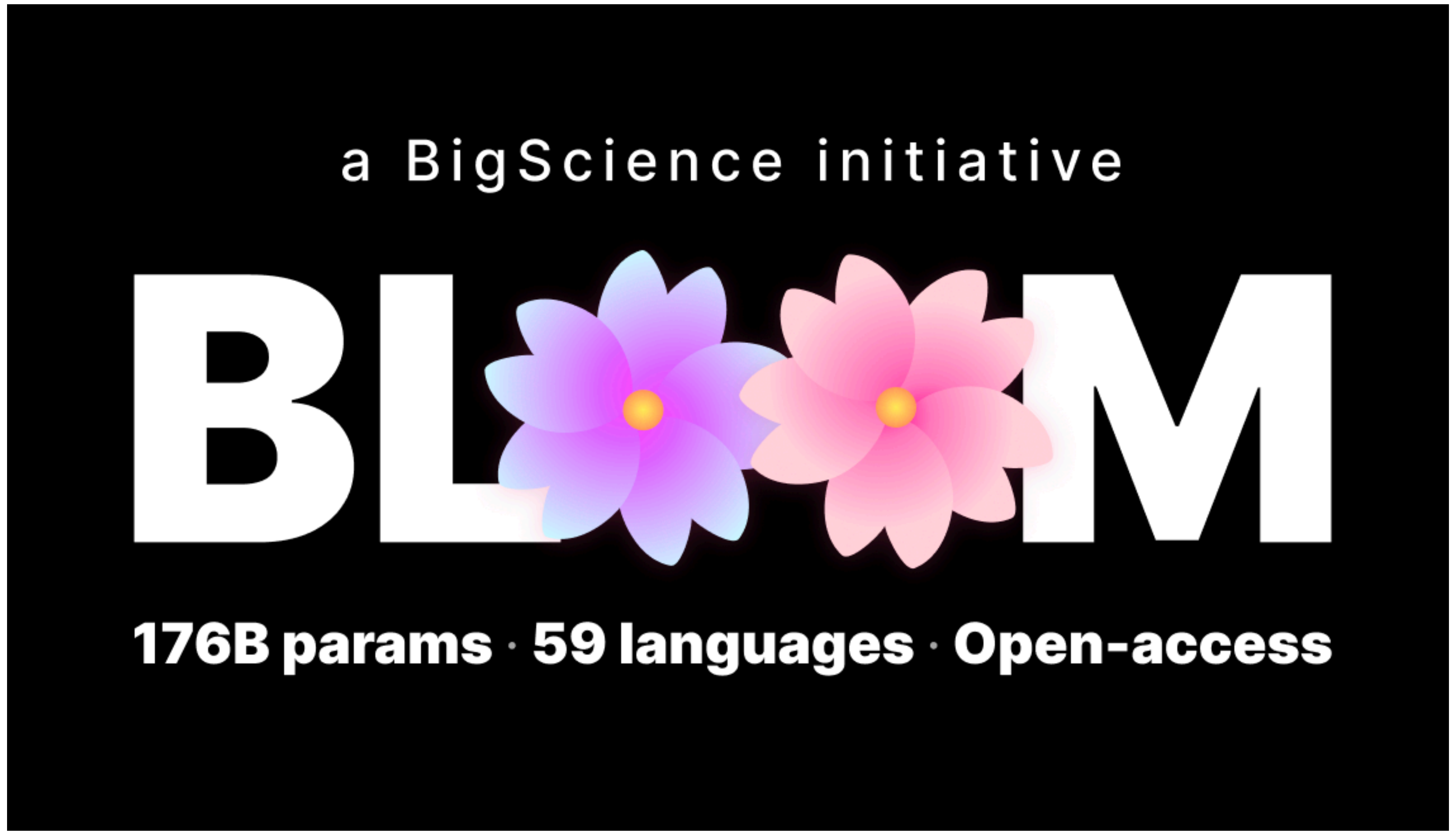
25+ ML libraries: Keras, spaCY, Scikit-Learn, fastai, etc.

10K organizations

100K+ users daily

1M+ downloads daily

The BLOOM Foundation Model



https://bigscience.huggingface.co

https://huggingface.co/bigscience/bloom

1.5TB of text, 350B tokens
43 languages, 16 programming languages

Estimating the carbon footprint of BLOOM

ESTIMATING THE CARBON FOOTPRINT OF BLOOM, A 176B PARAMETER LANGUAGE MODEL

Alexandra Sasha Luccioni
Hugging Face
sasha.luccioni@hf.co

Sylvain Viguier
Graphcore
sylvainv@graphcore.ai

Anne-Laure Ligozat
LISN & ENSIIE
anne-laure.ligozat@lisn.upsaclay.fr

ABSTRACT

Progress in machine learning (ML) comes with a cost to the environment, given that training ML models requires significant computational resources, energy and materials. In the present article, we aim to quantify the carbon footprint of BLOOM, a 176-billion parameter language model, across its life cycle. We estimate that BLOOM's final training emitted approximately 24.7 tonnes of CO_2eq if we consider only the dynamic power consumption, and 50.5 tonnes if we account for all processes ranging from equipment manufacturing to energy-based operational consumption. We also study the energy requirements and carbon emissions of its deployment for inference via an API endpoint receiving user queries in real-time. We conclude with a discussion regarding the difficulty of precisely estimating the carbon footprint of ML models and future research directions that can contribute towards improving carbon emissions reporting.

https://arxiv.org/abs/2211.02001

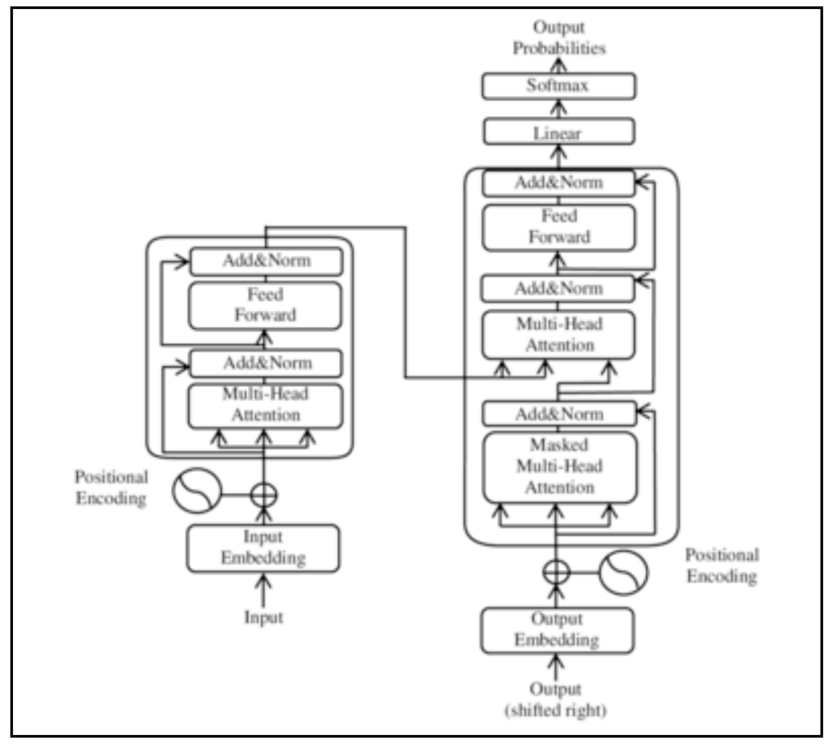
Training
118 days
384 A100 GPUs
124 A100-years
24.7 tons CO2eq (GPUs only)
50 tons CO2eq (total)

Inference
16 A100 GPUs
19 kgs CO2eq / day
7 tons CO2eq / year

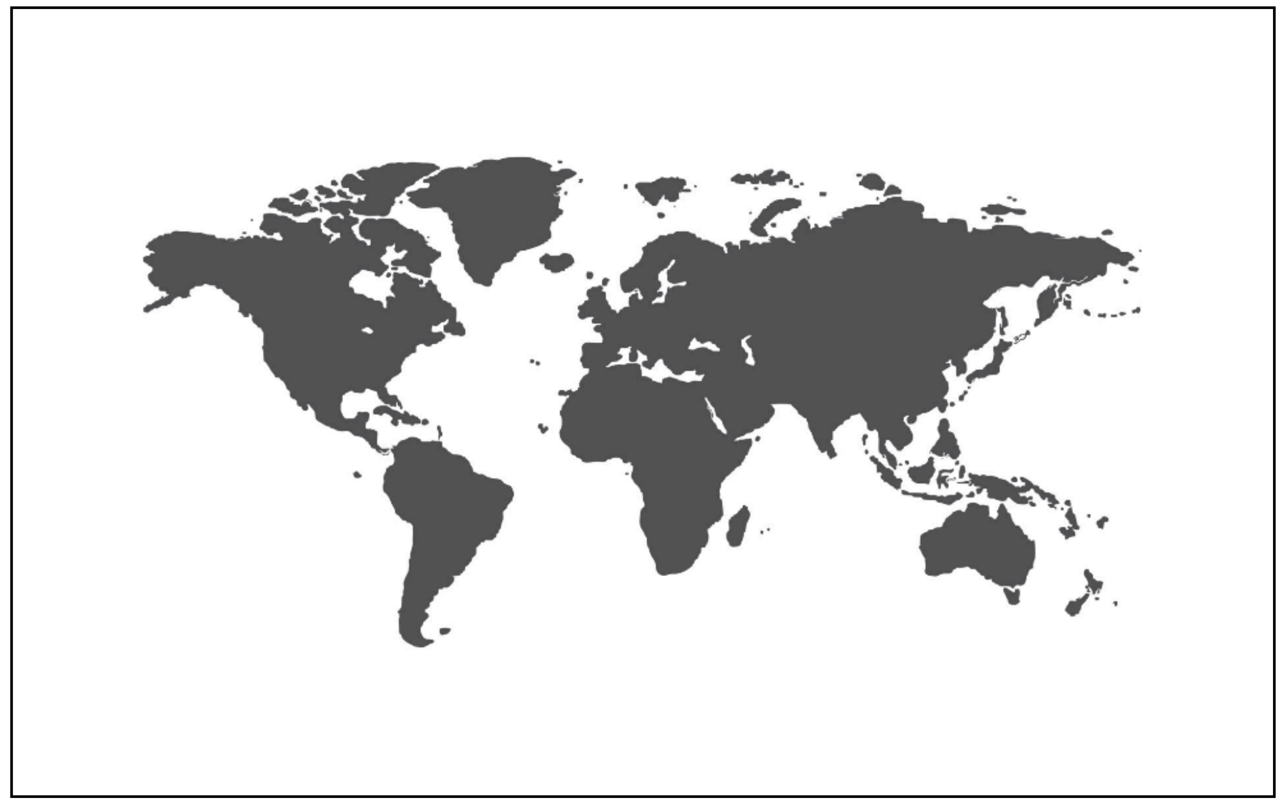
Reducing CO2 emissions for LLMs and FMs

Awareness

Model

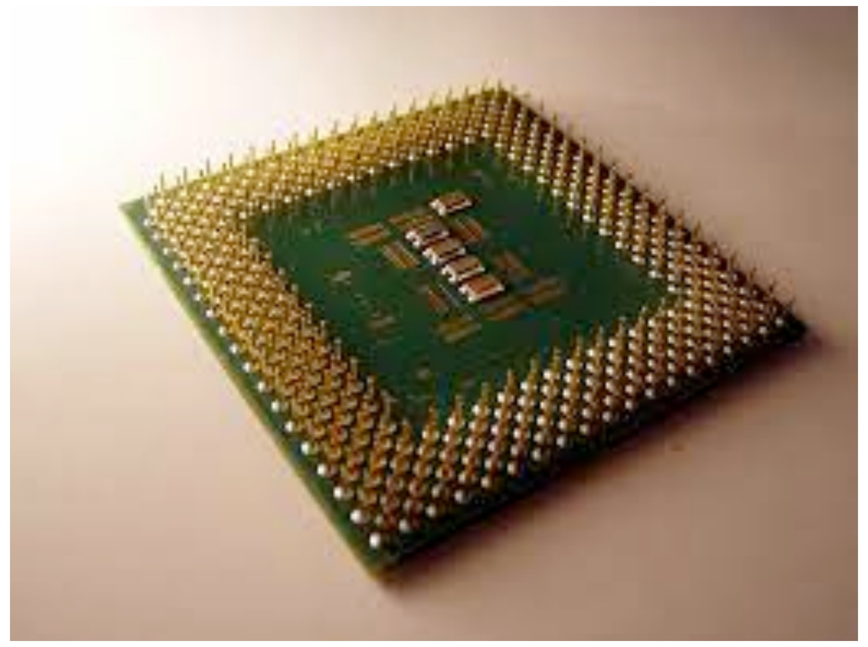


Power grid location



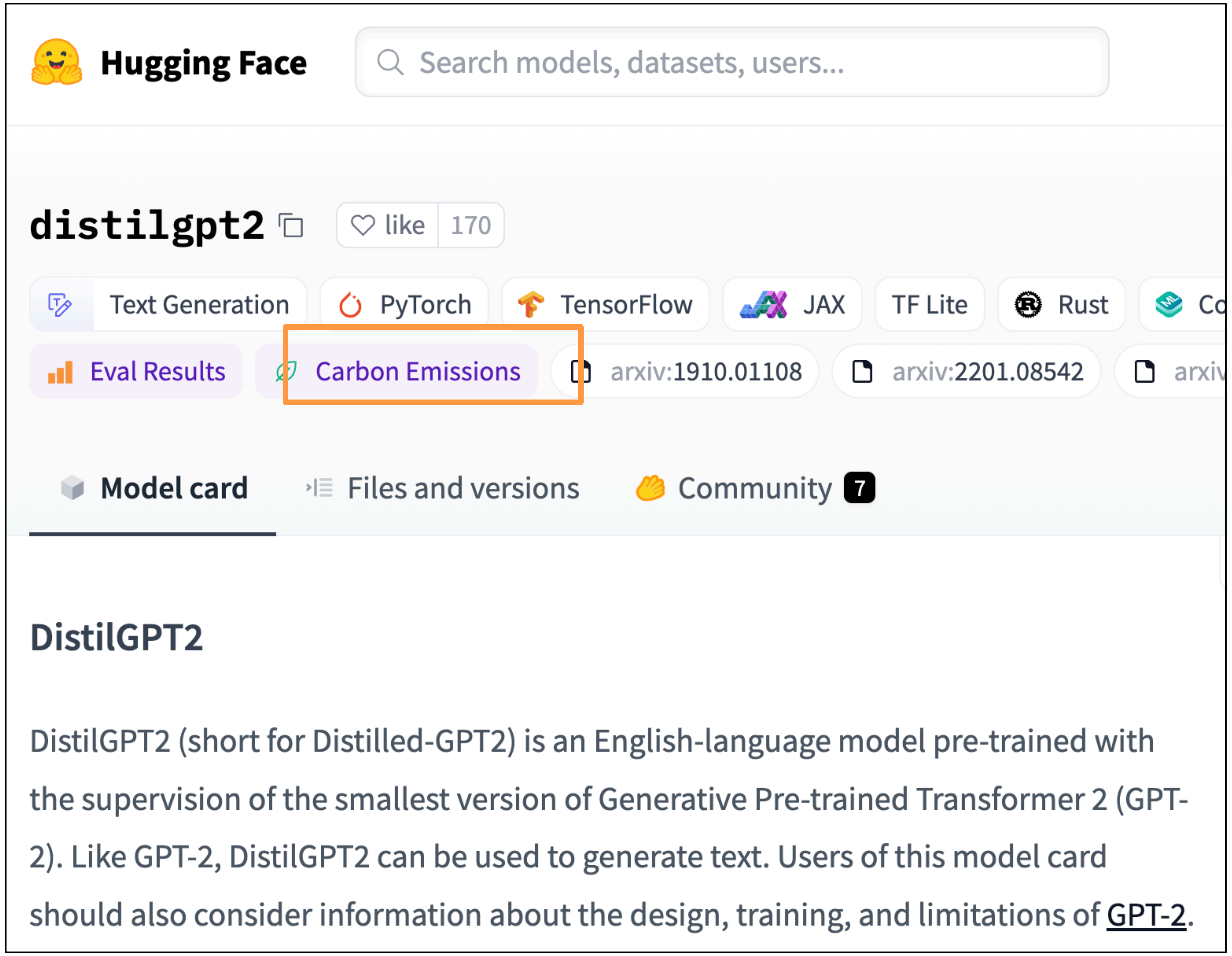
Infrastructure

Training and
inference hardware



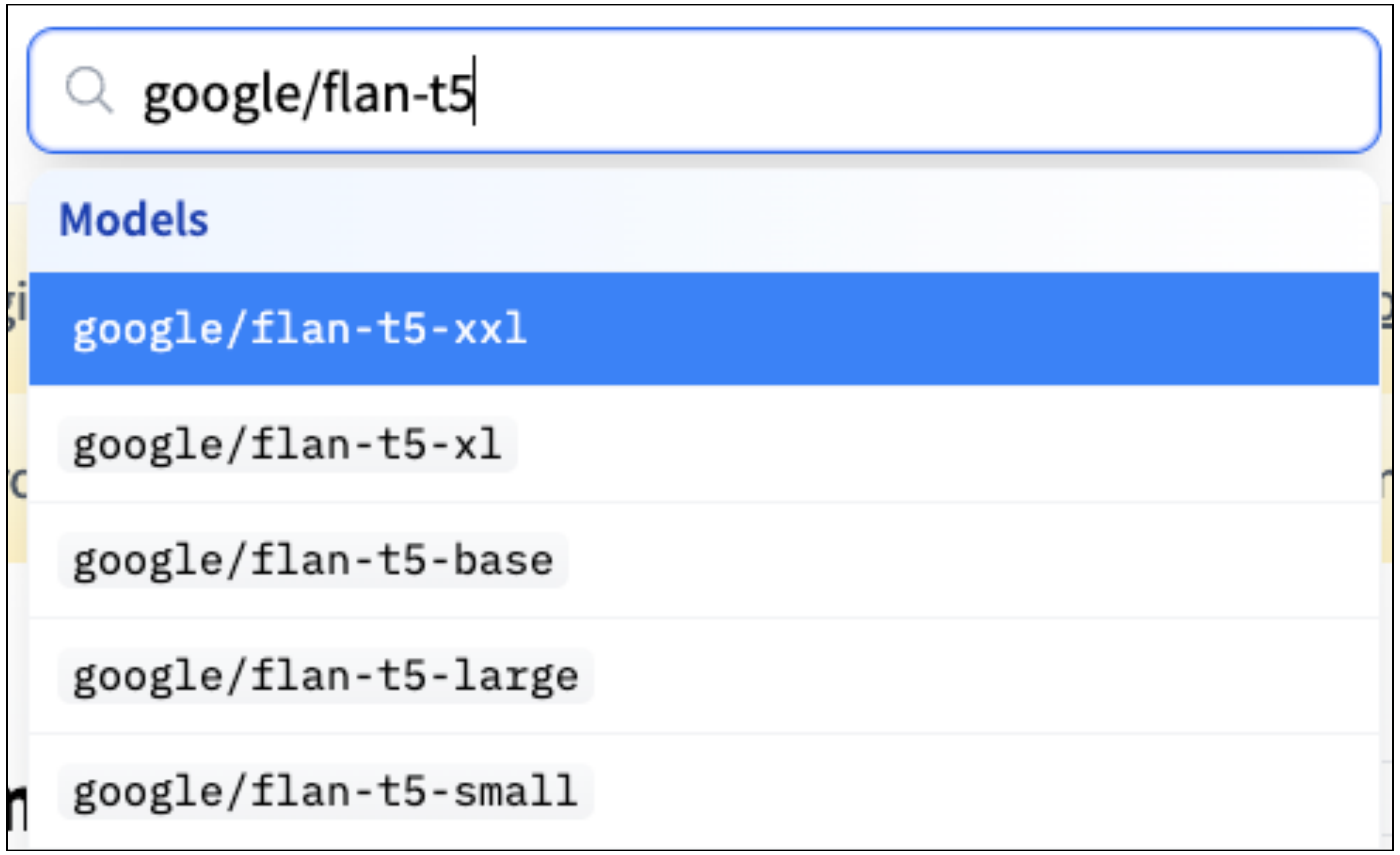
Increasing awareness

- **Energy efficiency** and **CO2 accounting** should be part of your project
- Your **ESG** team will need this information
- Some Hugging Face **model cards** already feature CO2 information
- You can automatically include it in your own models with **CodeCarbon**, see https://huggingface.co/blog/carbon-emissions-on-the-hub



Picking a model

- **Start small**, evaluate, and scale up if needed
- **Use off the shelf models** instead of inventing your own, particularly with NAS
- **Prefer fine-tuning** over initial training
- You probably **don't need HPO**, especially not grid search
- **Model optimization** goes a long way: FP16/BF16/FP8, INT quantization, pruning, etc.



Flan T5 models: from 80M to 11B parameters

Why customers are more successful with smaller models

- **Focus**: many business use cases require narrow domain knowledge
- **Agility**: they're faster to train and retrain, letting you iterate quicker
- **Cost**: they're much less expensive to train and host
- **Speed**: the smaller a model is, the faster it predicts
- **Accuracy**: a smaller model fine-tuned for a specific purpose will almost always outperform a larger general-purpose model

A selection of recent models

We introduce LLaMA, a collection of foundation language models ranging from 7B to 65B parameters. We train our models on trillions of tokens, and show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets. In particular, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B. We release all our models to the research community[1].

https://arxiv.org/abs/2302.13971

*We introduce **Alpaca 7B**, a model fine-tuned from the LLaMA 7B model on 52K instruction-following demonstrations. On our preliminary evaluation of single-turn instruction following, Alpaca behaves qualitatively similarly to OpenAI's text-davinci-003, while being surprisingly small and easy/cheap to reproduce (<600$).*

https://crfm.stanford.edu/2023/03/13/alpaca.html

information. With Multimodal-CoT, our model under 1 billion parameters outperforms the previous state-of-the-art LLM (GPT-3.5) by 16 percentage points (75.17%→91.68% accuracy) and even surpasses human performance on the ScienceQA benchmark. Code is publicly available.[1]

https://arxiv.org/abs/2302.00923

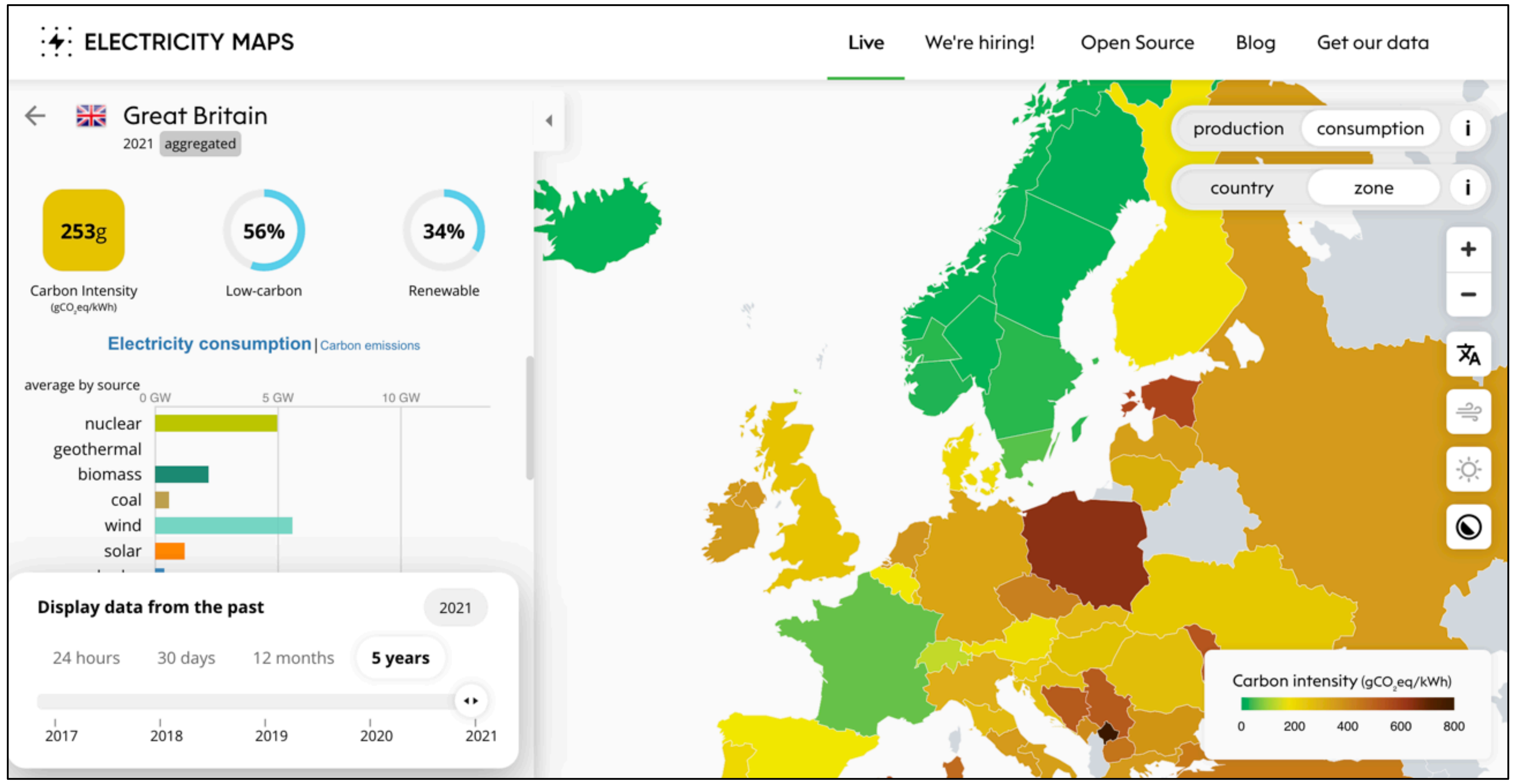
across multiple diverse setups. Finally, by scaling our model up to 20B parameters, we achieve SOTA performance on 50 well-established supervised NLP tasks ranging from language generation (with automated and human evaluation), language understanding, text classification, question answering, commonsense reasoning, long text reasoning, structured knowledge grounding and information retrieval. Our model also achieve strong results at in-context learning, outperforming 175B GPT-3 on zero-shot SuperGLUE and tripling the performance of T5-XXL on one-shot summarization.

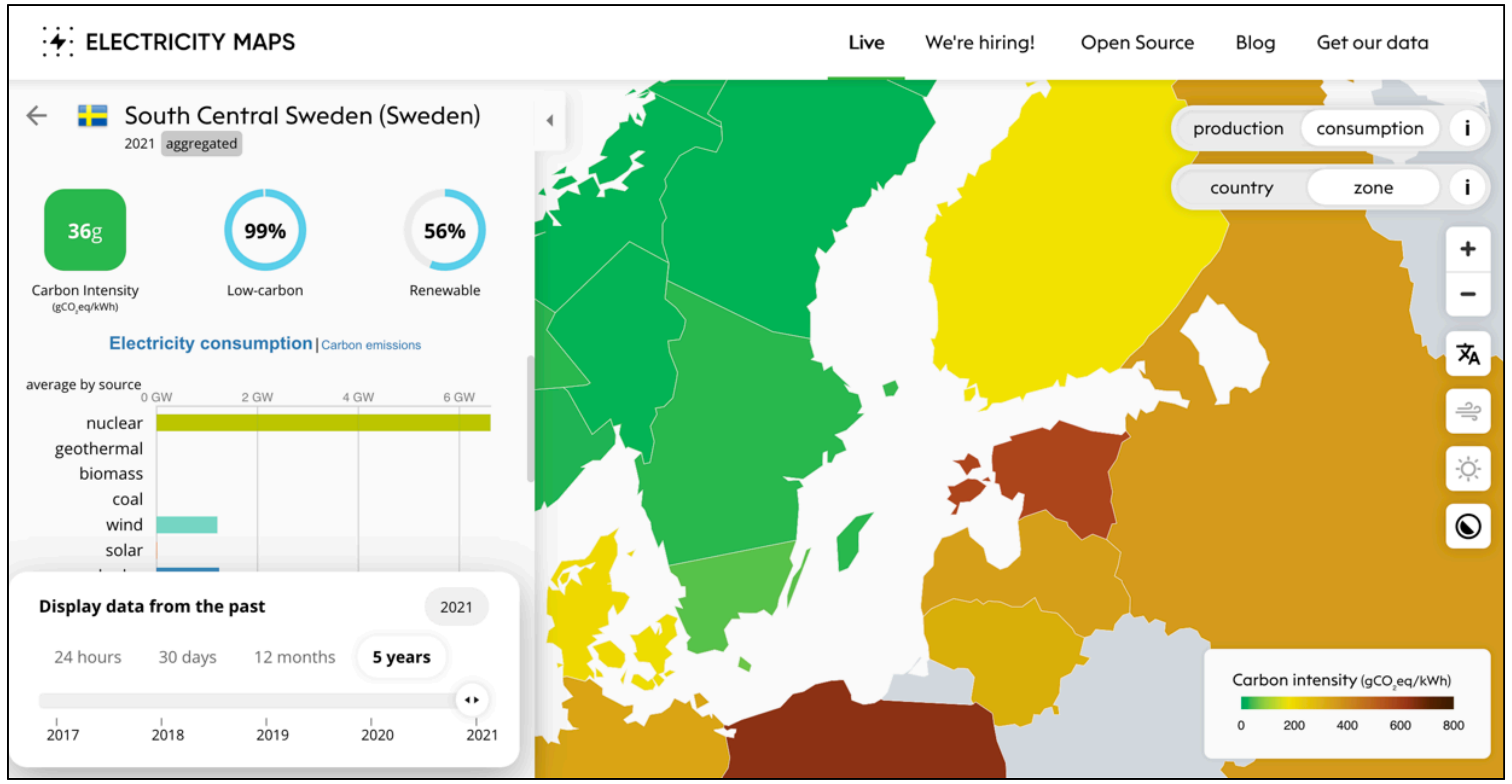
https://huggingface.co/google/flan-ul2

Picking a power grid location

- Power grids are **not equal** when it comes to CO2 emissions, even in Europe

- Using infrastructure hosted in a **greener** location will go a long way

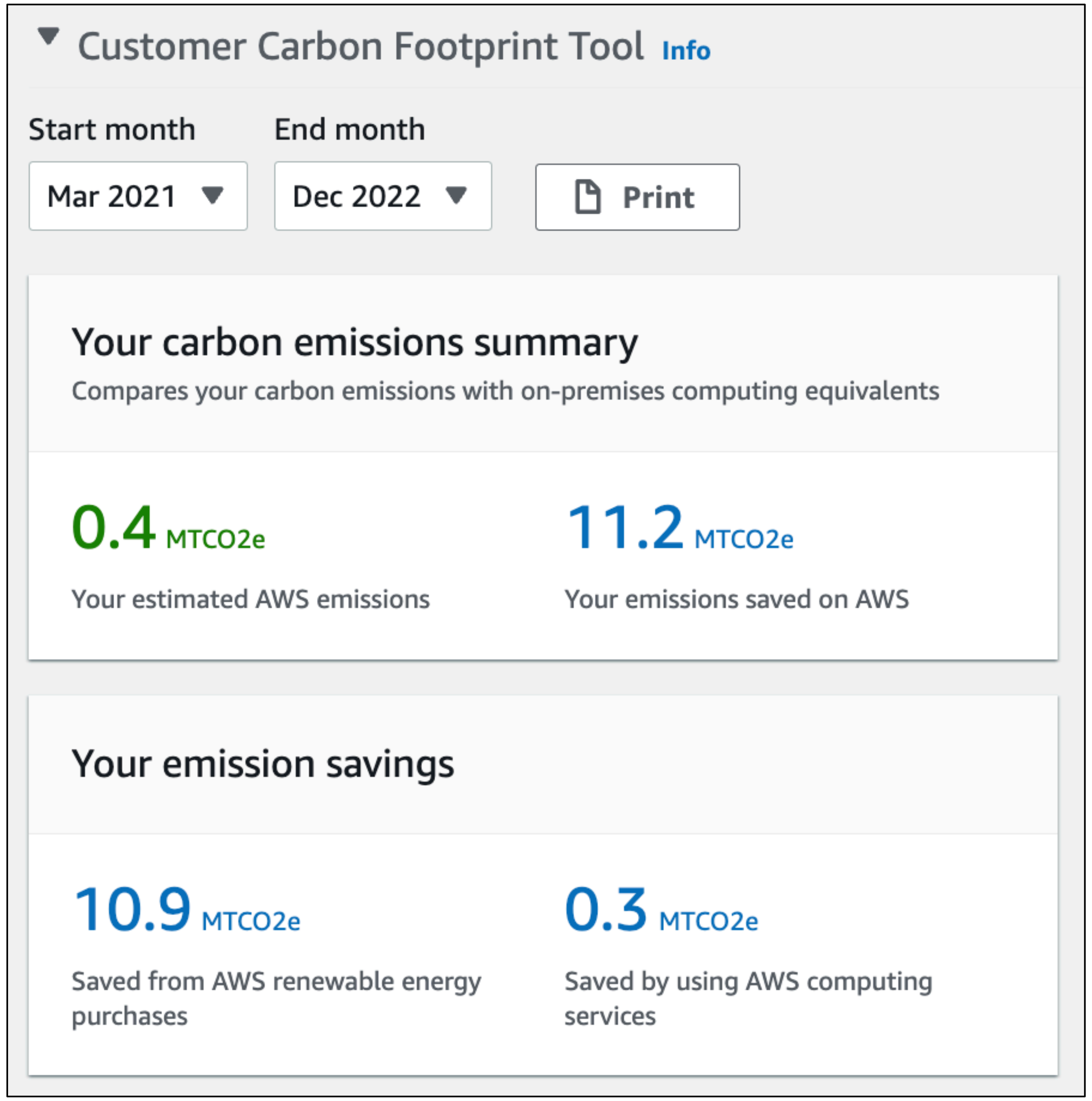




https://app.electricitymaps.com/map

Picking infrastructure

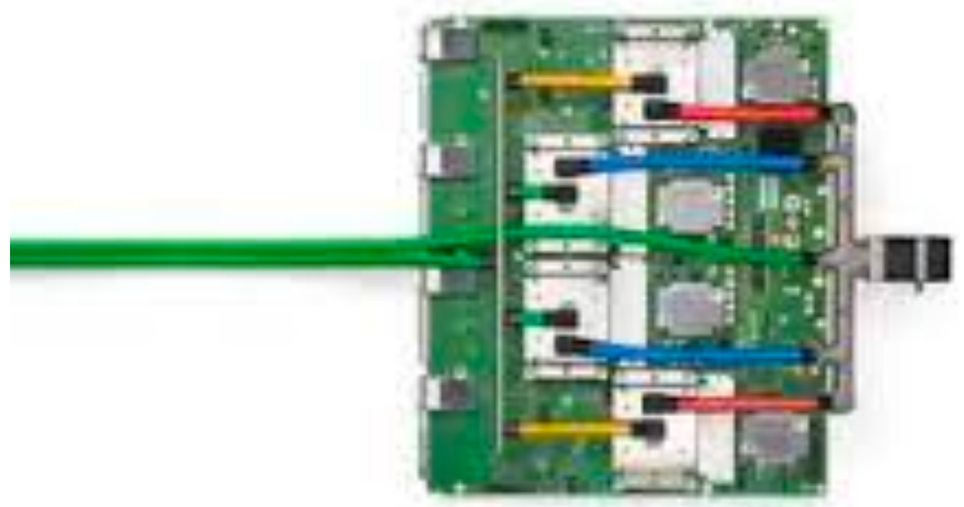
- **Cloud infrastructure** consistently provides better efficiency than on-premises infrastructure
  - Economies of scale, deep expertise, on-demand vs. always-on, etc.
- For example, Amazon is world's largest corporate purchaser of **renewable energy**
  - AWS: **3.6x** more efficient than the median of US enterprise DCs, and up to **5x** when compared to European DCs
  - AWS: **80%** reduction in carbon footprint compared to enterprise DCs
  - On track to power **100%** of their operations with renewable energy by 2025
  - https://aws.amazon.com/energy/sustainability/
- You can easily find the **greenest cloud region**
  - https://mlco2.github.io/impact/



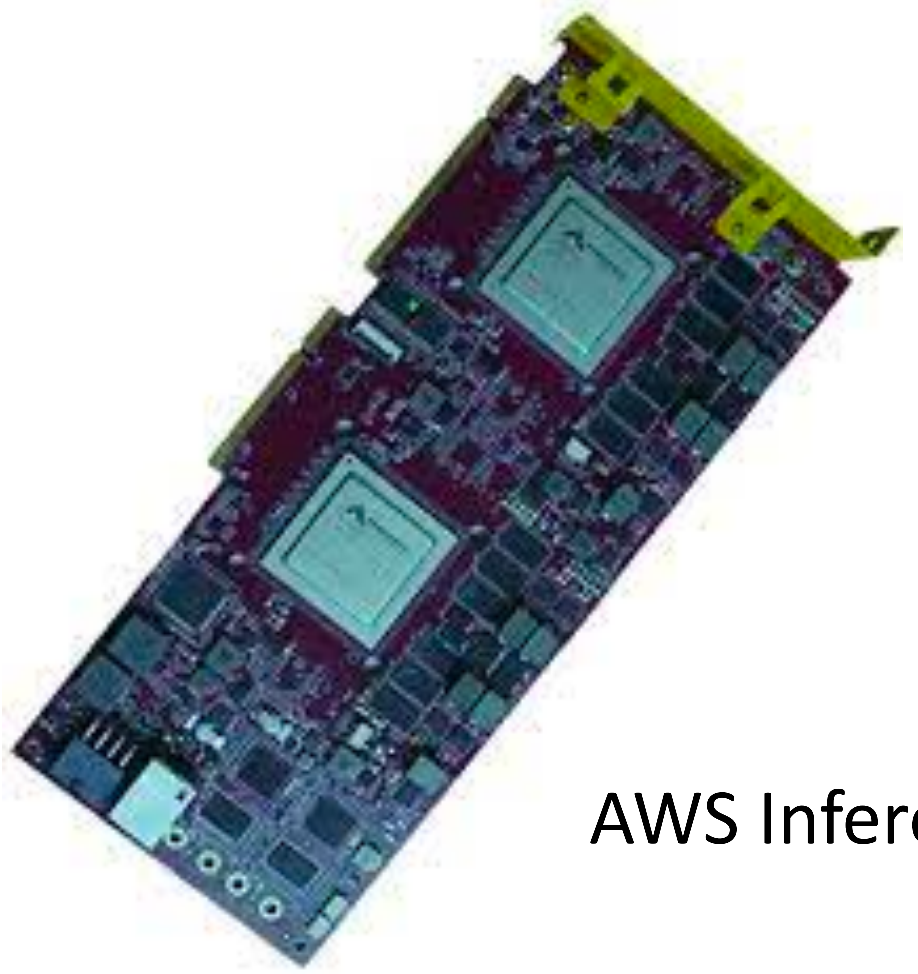
Picking training and inference hardware

- There's more to life than (larger and larger) GPUs
  - Your **latency budget** may not require a GPU
  - **CPU inference** works great for many NLP workloads
  - Stable Diffusion on CPU in under **5 seconds** https://www.youtube.com/watch?v=KJDCGyZ2fPw
  - **CPU optimization tools**: Intel Neural Compressor, Intel OpenVINO, Hugging Face Optimum Intel

- AI accelerators
  - **Google TPUv4**: 1.2x–1.7x faster and uses 1.3x–1.9x less power than the A100
  - **AWS Trainium**, **AWS Inferentia**: better cost-performance and better performance-per-watt compared to GPUs
  - **Intel Habana Gaudi 2**: 3x faster inference for BLOOMZ-7B than the A100 https://huggingface.co/blog/habana-gaudi-2-bloom
  - **Not difficult to switch**: Hugging Face Accelerate, Hugging Face Optimum libraries



TPU v4



AWS Inferentia

90% of Google models are trained on TPU

DNN Model	*TPU v1 7/2016 (Inference)*	*TPU v3 4/2019 (Training & Inference)*	*TPU v4 Lite 2/2020 (Inference)*	*TPU v4 10/2022 (Training)*
MLP/DLRM	61%	27%	25%	24%
RNN	29%	21%	29%	2%
CNN	5%	24%	18%	12%
Transformer	--	21%	28%	57%
(BERT)	--	--	*(28%)*	*(26%)*
(LLM)	--	--	--	*(31%)*

https://arxiv.org/abs/2304.01433

Summing things up

- Transformer models are the de facto standard for AI-powered apps.
- Training and deploying these large models have an environmental impact
- We can significantly reduce the impact with a combination of:
  - Small, pre-trained models fine-tuned on domain-specific data,
  - Efficient cloud infrastructure,
  - Cost and power-efficient AI hardware

Getting started

Stay in touch!

@julsimon
julsimon.medium.com
youtube.com/c/juliensimonfr

https://huggingface.co/tasks

https://huggingface.co/course

https://github.com/huggingface

https://huggingface.co/blog