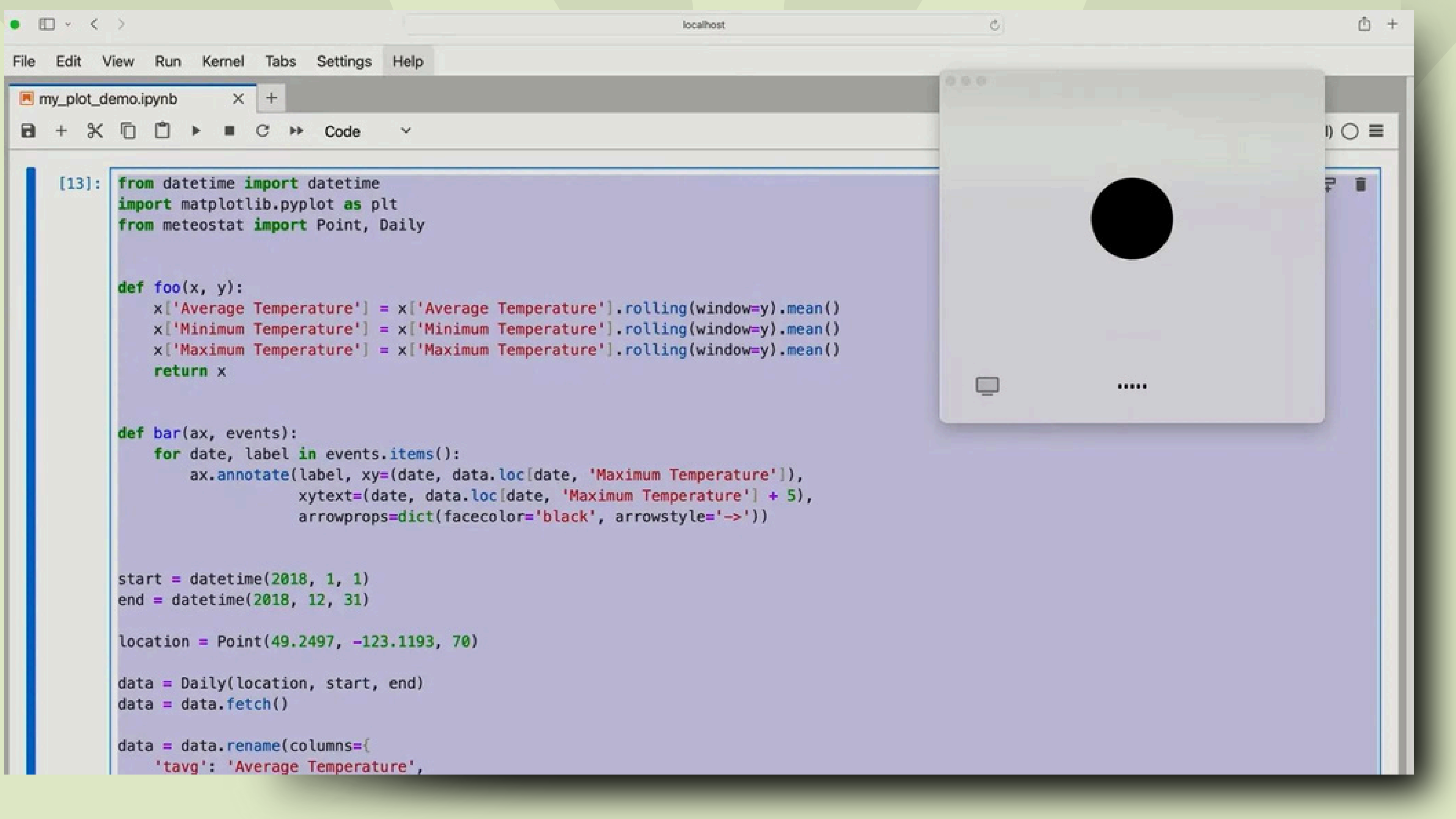
Dr. Martha Boeckenfeld

14 BEST LARGE LANGUAGE MODELS (LLM) IN 2024

Swipe →

1. GPT-4 O (OMNI)

OpenAI is launching GPT-4o, an iteration of the GPT-4 model that powers its hallmark product, ChatGPT. The updated model “is much faster” and improves “capabilities across text, vision, and audio,” OpenAI CTO Mira Murati said in a livestream announcement. It’ll be free for all users, and paid users will continue to “have up to five times the capacity limits” of free users, Murati added.



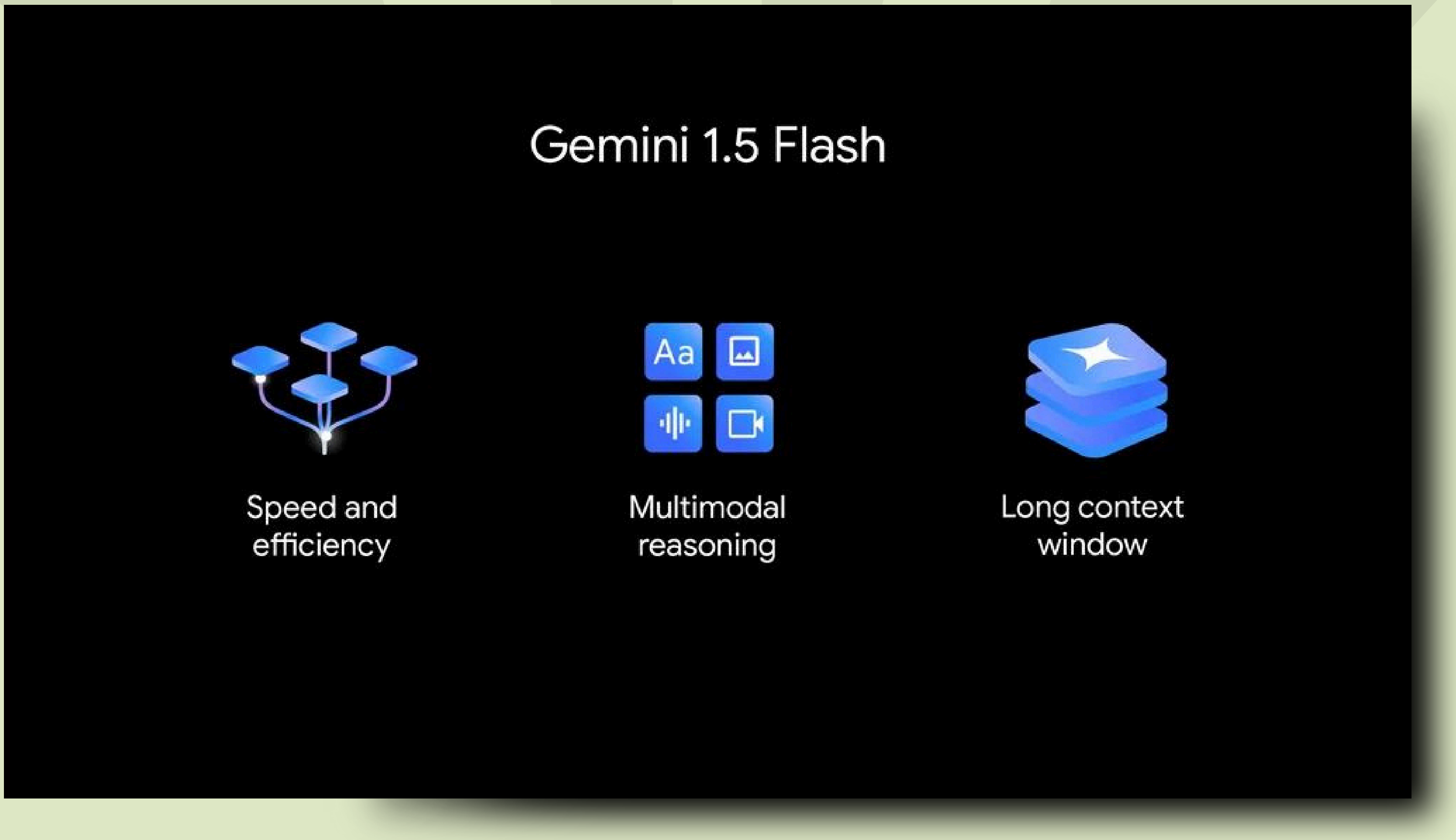
```
[13]: from datetime import datetime
      import matplotlib.pyplot as plt
      from meteostat import Point, Daily


      def foo(x, y):
          x['Average Temperature'] = x['Average Temperature'].rolling(window=y).mean()
          x['Minimum Temperature'] = x['Minimum Temperature'].rolling(window=y).mean()
          x['Maximum Temperature'] = x['Maximum Temperature'].rolling(window=y).mean()
          return x


      def bar(ax, events):
          for date, label in events.items():
              ax.annotate(label, xy=(date, data.loc[date, 'Maximum Temperature']),
                          xytext=(date, data.loc[date, 'Maximum Temperature'] + 5),
                          arrowprops=dict(facecolor='black', arrowstyle='->'))


      start = datetime(2018, 1, 1)
      end = datetime(2018, 12, 31)

      location = Point(49.2497, -123.1193, 70)

      data = Daily(location, start, end)
      data = data.fetch()

      data = data.rename(columns={
          'tavg': 'Average Temperature',
```

PREV

NEXT

2. GEMINI FLASH 1.5

Google has also introduced a new addition to the family of Gemini AI models. The new AI model, dubbed Gemini 1.5 Flash is a light-weight model that is designed to be faster, more responsive, and cost-efficient. The tech giant said that it has worked on improving its latency to improve its speed. While solving complex tasks would not be its strength, it can do tasks such as summarisation, chat applications, image and video captioning, data extraction from long documents and tables, and more.



Gemini 1.5 Flash

Speed and efficiency

Multimodal reasoning

Long context window

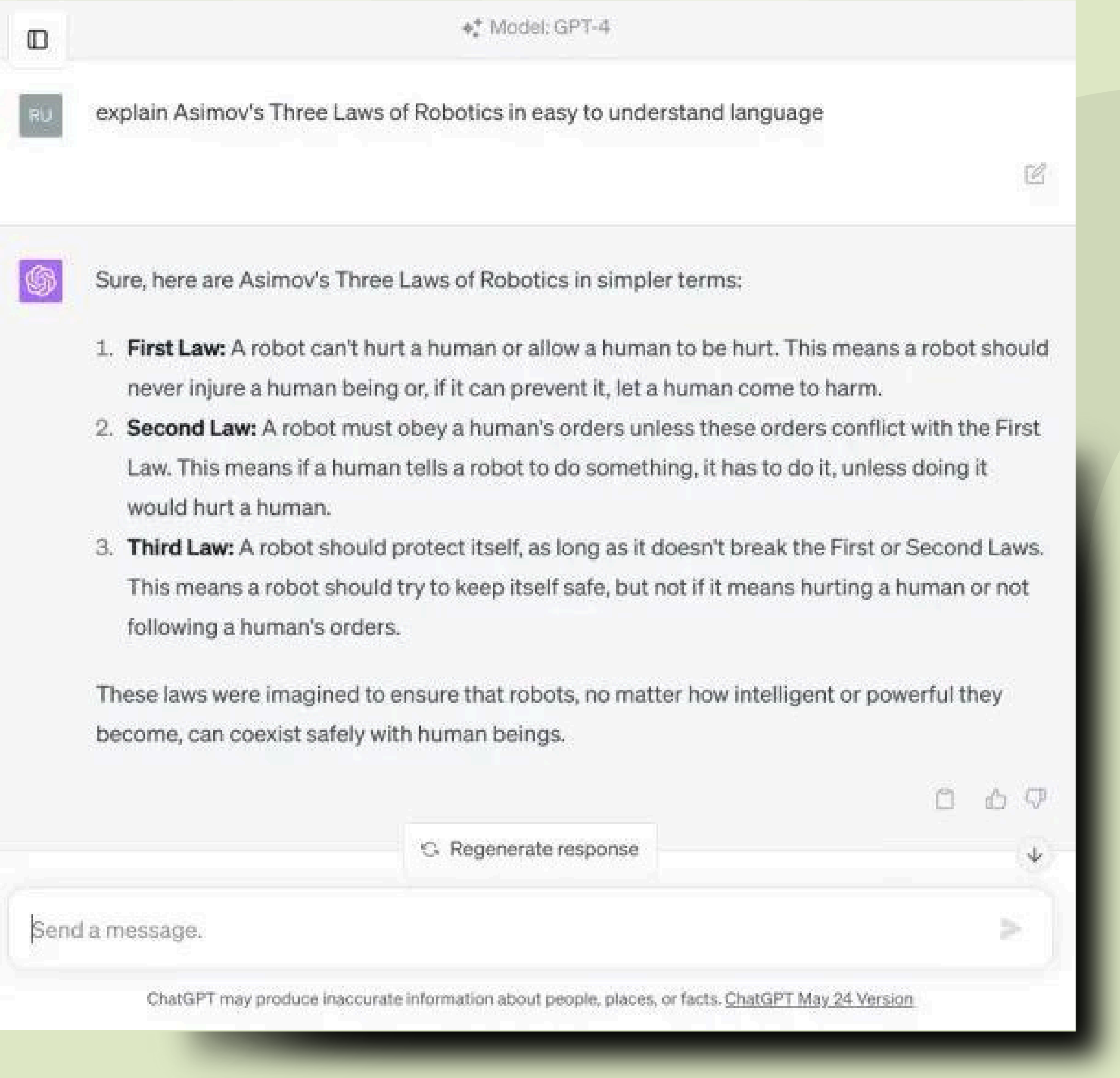
PREV

NEXT

Dr. Martha Boeckenfeld

3. GPT-4

The GPT-4 model by OpenAI is the best AI large language model (LLM) available in 2024. Released in March 2023, the GPT-4 model has showcased tremendous capabilities with complex reasoning understanding, advanced coding capability, proficiency in multiple academic exams, skills that exhibit human-level performance, and much more



explain Asimov's Three Laws of Robotics in easy to understand language

Sure, here are Asimov's Three Laws of Robotics in simpler terms:

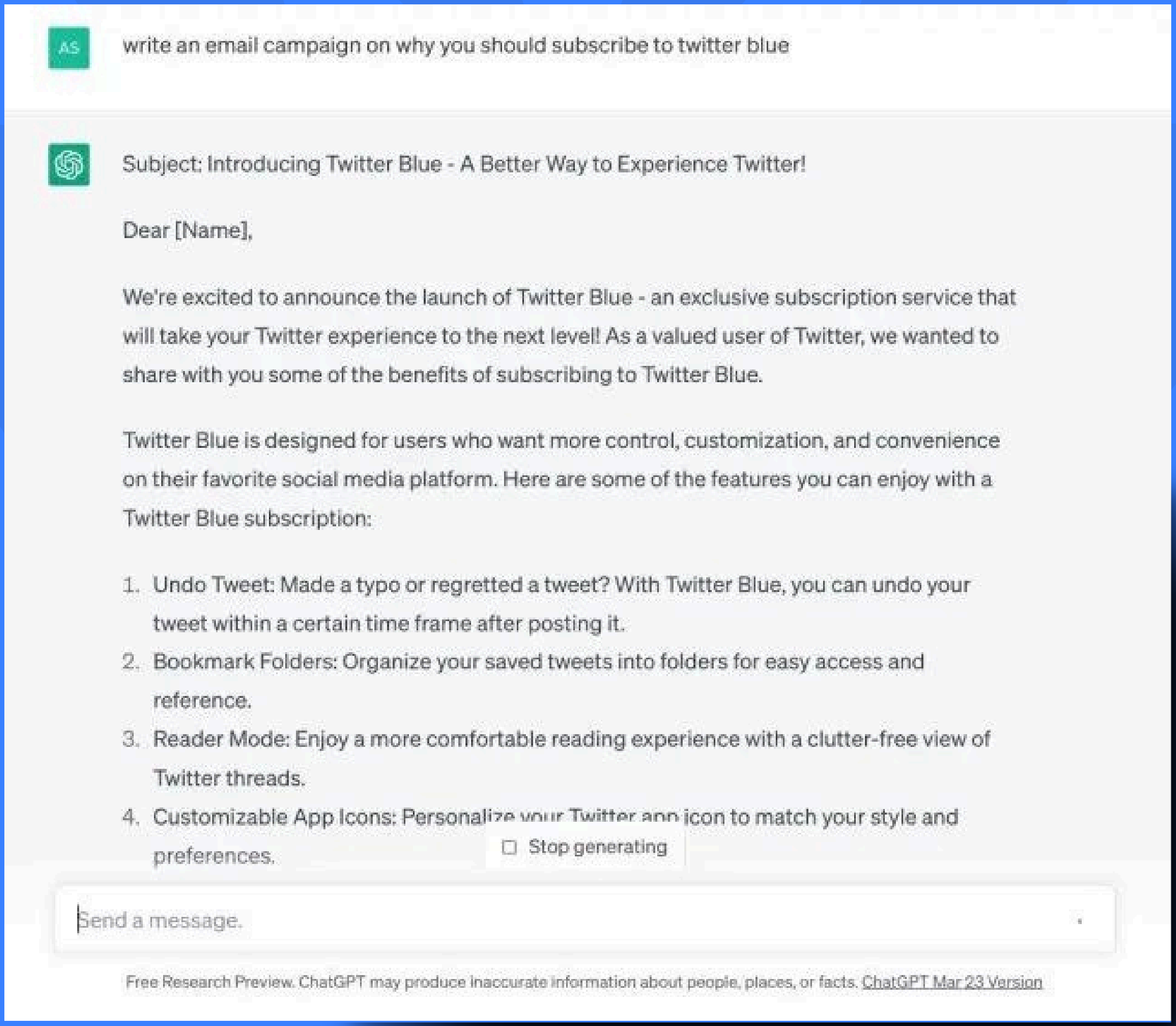
1. **First Law:** A robot can't hurt a human or allow a human to be hurt. This means a robot should never injure a human being or, if it can prevent it, let a human come to harm.
2. **Second Law:** A robot must obey a human's orders unless these orders conflict with the First Law. This means if a human tells a robot to do something, it has to do it, unless doing it would hurt a human.
3. **Third Law:** A robot should protect itself, as long as it doesn't break the First or Second Laws. This means a robot should try to keep itself safe, but not if it means hurting a human or not following a human's orders.

These laws were imagined to ensure that robots, no matter how intelligent or powerful they become, can coexist safely with human beings.

ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT May 24 Version

4. GPT-3.5

After GPT 4, OpenAI takes the second spot again with GPT-3.5. It's a general-purpose LLM similar to GPT-4 but lacks expertise in specific domains. Talking about the pros first, it's an incredibly fast model and generates a complete response within seconds.



write an email campaign on why you should subscribe to twitter blue

Subject: Introducing Twitter Blue - A Better Way to Experience Twitter!

Dear [Name],

We're excited to announce the launch of Twitter Blue - an exclusive subscription service that will take your Twitter experience to the next level! As a valued user of Twitter, we wanted to share with you some of the benefits of subscribing to Twitter Blue.

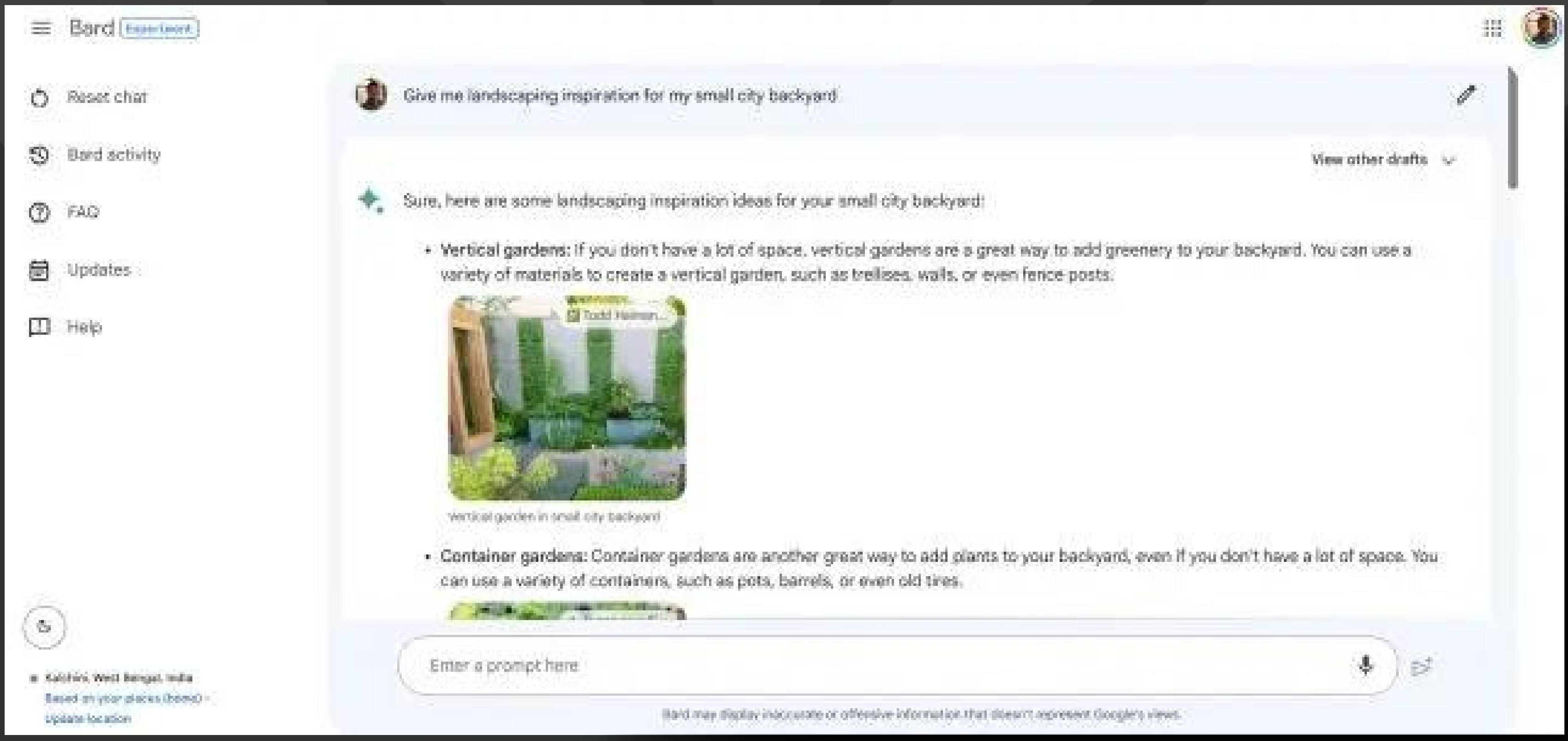
Twitter Blue is designed for users who want more control, customization, and convenience on their favorite social media platform. Here are some of the features you can enjoy with a Twitter Blue subscription:

1. Undo Tweet: Made a typo or regretted a tweet? With Twitter Blue, you can undo your tweet within a certain time frame after posting it.
2. Bookmark Folders: Organize your saved tweets into folders for easy access and reference.
3. Reader Mode: Enjoy a more comfortable reading experience with a clutter-free view of Twitter threads.
4. Customizable App Icons: Personalize your Twitter app icon to match your style and preferences.

Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT Mar 23 Version

5. PALM 2 (BISON-001)

Next, we have the PaLM 2 AI model from Google, which is ranked among the best large language models of 2024. Google has focused on commonsense reasoning, formal logic, mathematics, and advanced coding in 20+ languages on the PaLM 2 model. It’s being said that the largest PaLM 2 model has been trained on 540 billion parameters and has a maximum context length of 4096 tokens.


Bard
Experiment
Reset chat
Bard activity
FAQ
Updates
Help
View other drafts
Todd Heiman...
Vertical garden in small city backyard
Enter a prompt here
Kalchini, West Bengal, India
Based on your places (home) · Update location

Give me landscaping inspiration for my small city backyard

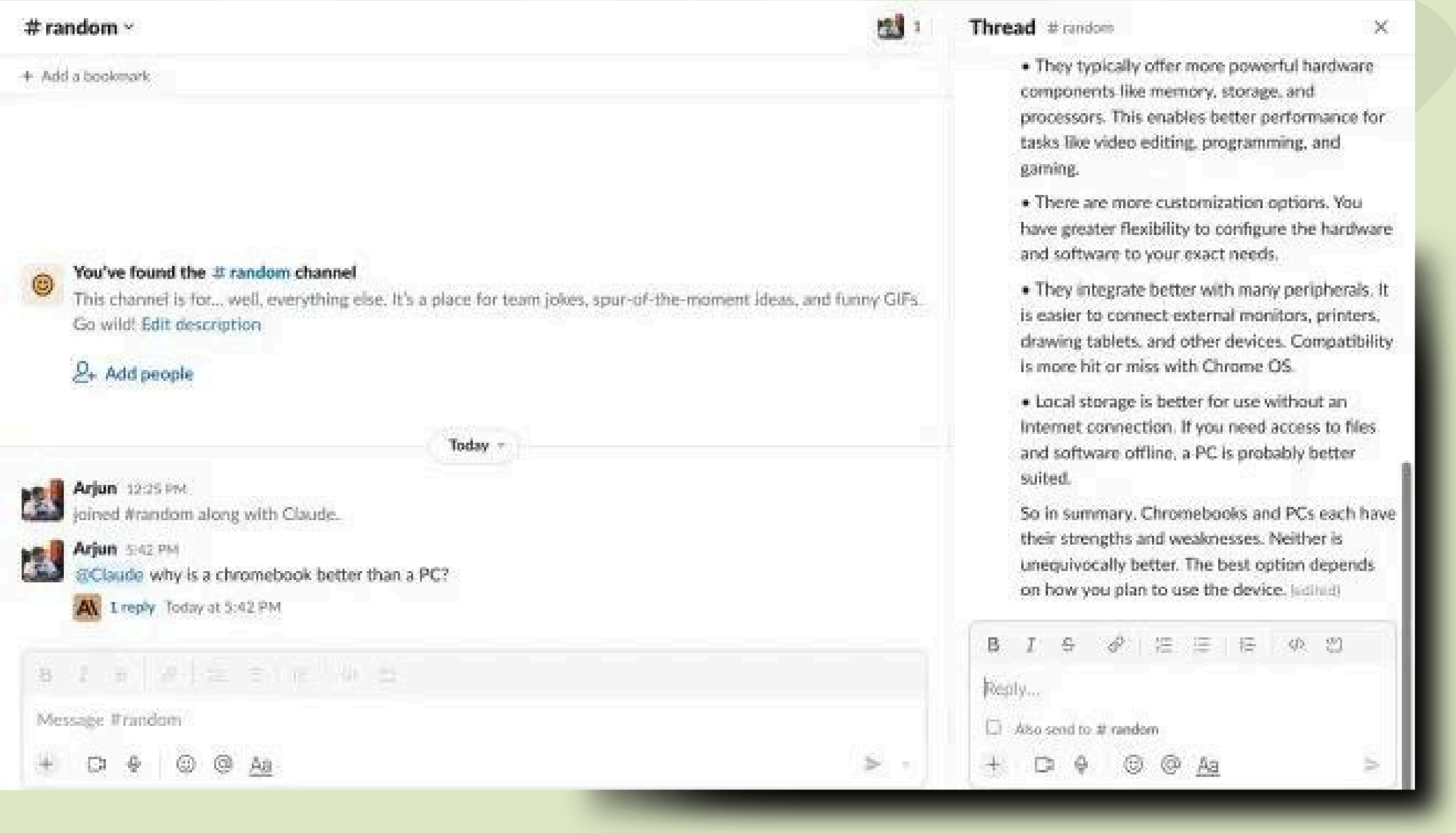
Sure, here are some landscaping inspiration ideas for your small city backyard:

- **Vertical gardens:** If you don't have a lot of space, vertical gardens are a great way to add greenery to your backyard. You can use a variety of materials to create a vertical garden, such as trellises, walls, or even fence posts.
- **Container gardens:** Container gardens are another great way to add plants to your backyard, even if you don't have a lot of space. You can use a variety of containers, such as pots, barrels, or even old tires.

Bard may display inaccurate or offensive information that doesn't represent Google's views.

6. CLAUDE V1

In case you are unaware, Claude is a powerful LLM developed by Anthropic, which has been backed by Google. It has been co-founded by former OpenAI employees and its approach is to build AI assistants which are helpful, honest, and harmless. In multiple benchmark tests, Anthropic's Claude v1 and Claude Instant models have shown great promise. In fact, Claude v1 performs better than PaLM 2 in MMLU and MT-Bench tests.



You've found the # random channel

This channel is for... well, everything else. It's a place for team jokes, spur-of-the-moment ideas, and funny GIFs. Go wild! Edit description

Add people

Arjun 12:25 PM
joined #random along with Claude.

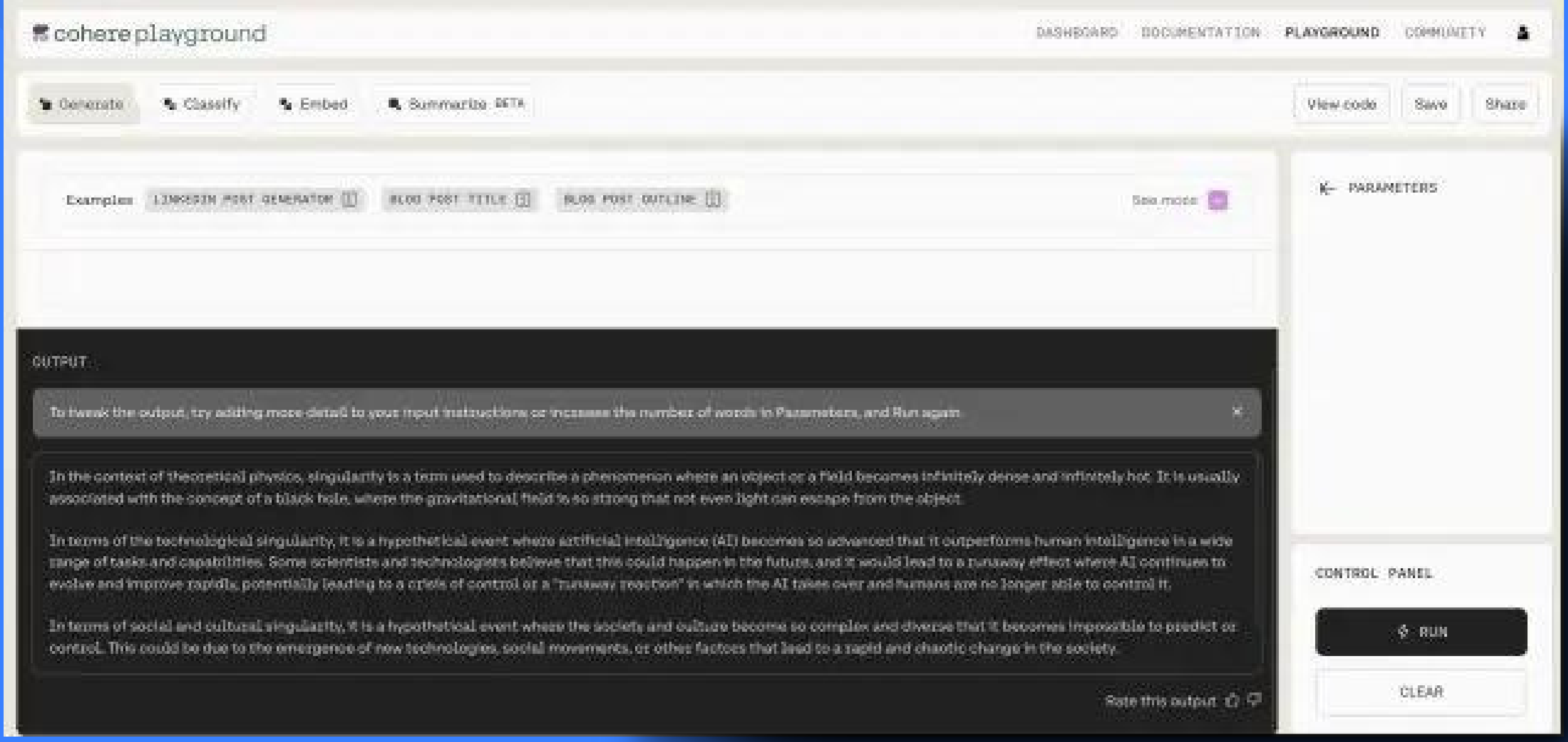
Arjun 5:42 PM
@Claude why is a chromebook better than a PC?
1 reply Today at 5:42 PM

- They typically offer more powerful hardware components like memory, storage, and processors. This enables better performance for tasks like video editing, programming, and gaming.
- There are more customization options. You have greater flexibility to configure the hardware and software to your exact needs.
- They integrate better with many peripherals. It is easier to connect external monitors, printers, drawing tablets, and other devices. Compatibility is more hit or miss with Chrome OS.
- Local storage is better for use without an Internet connection. If you need access to files and software offline, a PC is probably better suited.

So in summary, Chromebooks and PCs each have their strengths and weaknesses. Neither is unequivocally better. The best option depends on how you plan to use the device. (edited)

7. COHERE

Cohere is an AI startup founded by former Google employees who worked on the Google Brain team. One of its co-founders, Aidan Gomez was part of the "Attention is all you Need" paper that introduced the Transformer architecture. Unlike other AI companies, Cohere is here for enterprises and solving generative AI use cases for corporations. Cohere has a number of models from small to large – having just 6B parameters to large models trained on 52B parameters.



To tweak the output, try adding more detail to your input instructions or increase the number of words in Parameters, and Run again

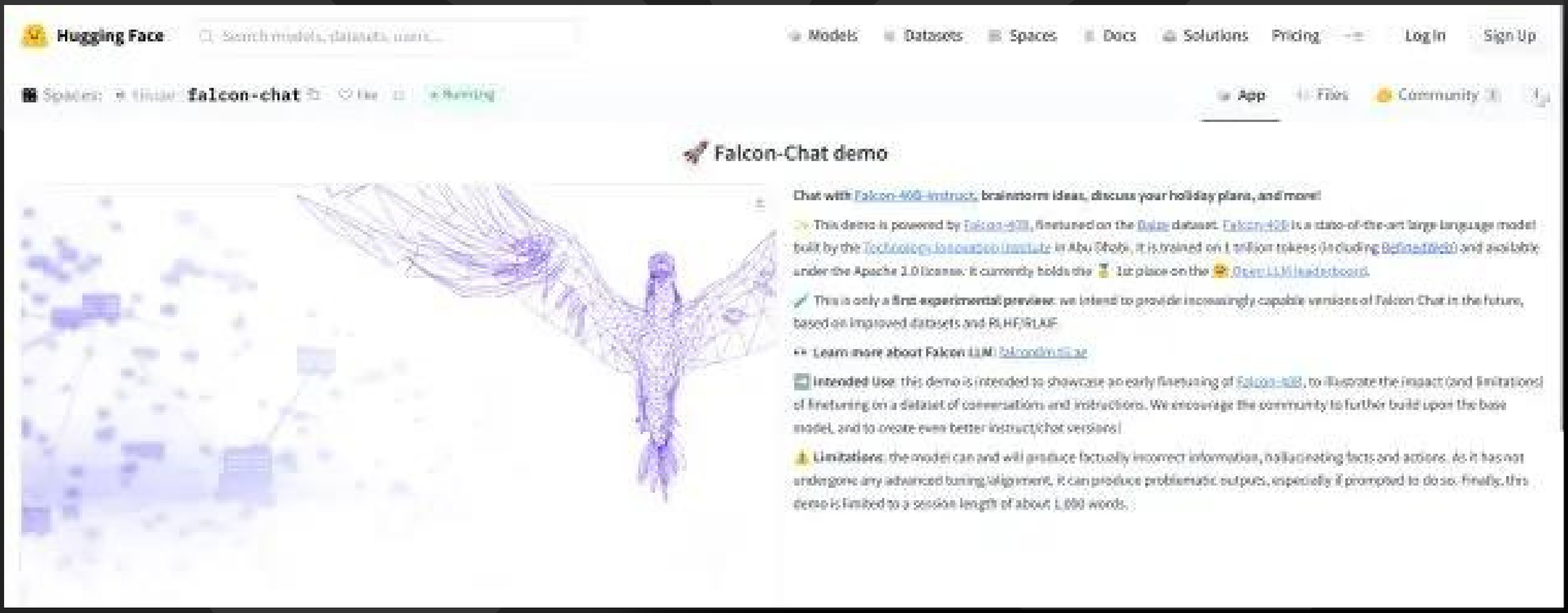
In the context of theoretical physics, singularity is a term used to describe a phenomenon where an object or a field becomes infinitely dense and infinitely hot. It is usually associated with the concept of a black hole, where the gravitational field is so strong that not even light can escape from the object.

In terms of the technological singularity, it is a hypothetical event where artificial intelligence (AI) becomes so advanced that it outperforms human intelligence in a wide range of tasks and capabilities. Some scientists and technologists believe that this could happen in the future, and it would lead to a runaway effect where AI continues to evolve and improve rapidly, potentially leading to a crisis of control or a "runaway reaction" in which the AI takes over and humans are no longer able to control it.

In terms of social and cultural singularity, it is a hypothetical event where the society and culture become so complex and diverse that it becomes impossible to predict or control. This could be due to the emergence of new technologies, social movements, or other factors that lead to a rapid and chaotic change in the society.

8. FALCON

Falcon is the first open-source large language model on this list, and it has outranked all the open-source models released so far, including LLaMA, StableLM, MPT, and more. It has been developed by the Technology Innovation Institute (TII), UAE. The best thing about Falcon is that it has been open-sourced with Apache 2.0 license, which means you can use the model for commercial purposes. There are no royalties or restrictions either.



🚀 Falcon-Chat demo

Chat with Falcon-40B-instruct, brainstorm ideas, discuss your holiday plans, and more!

This demo is powered by Falcon-40B, finetuned on the Baize dataset. Falcon-40B is a state-of-the-art large language model built by the Technology Innovation Institute in Abu Dhabi. It is trained on 1 trillion tokens (including RefinedWeb) and available under the Apache 2.0 license. It currently holds the 🥇 1st place on the 🤗 Open LLM leaderboard.

✓ This is only a **first experimental preview**: we intend to provide increasingly capable versions of Falcon Chat in the future, based on improved datasets and RLHF/RLAIF.

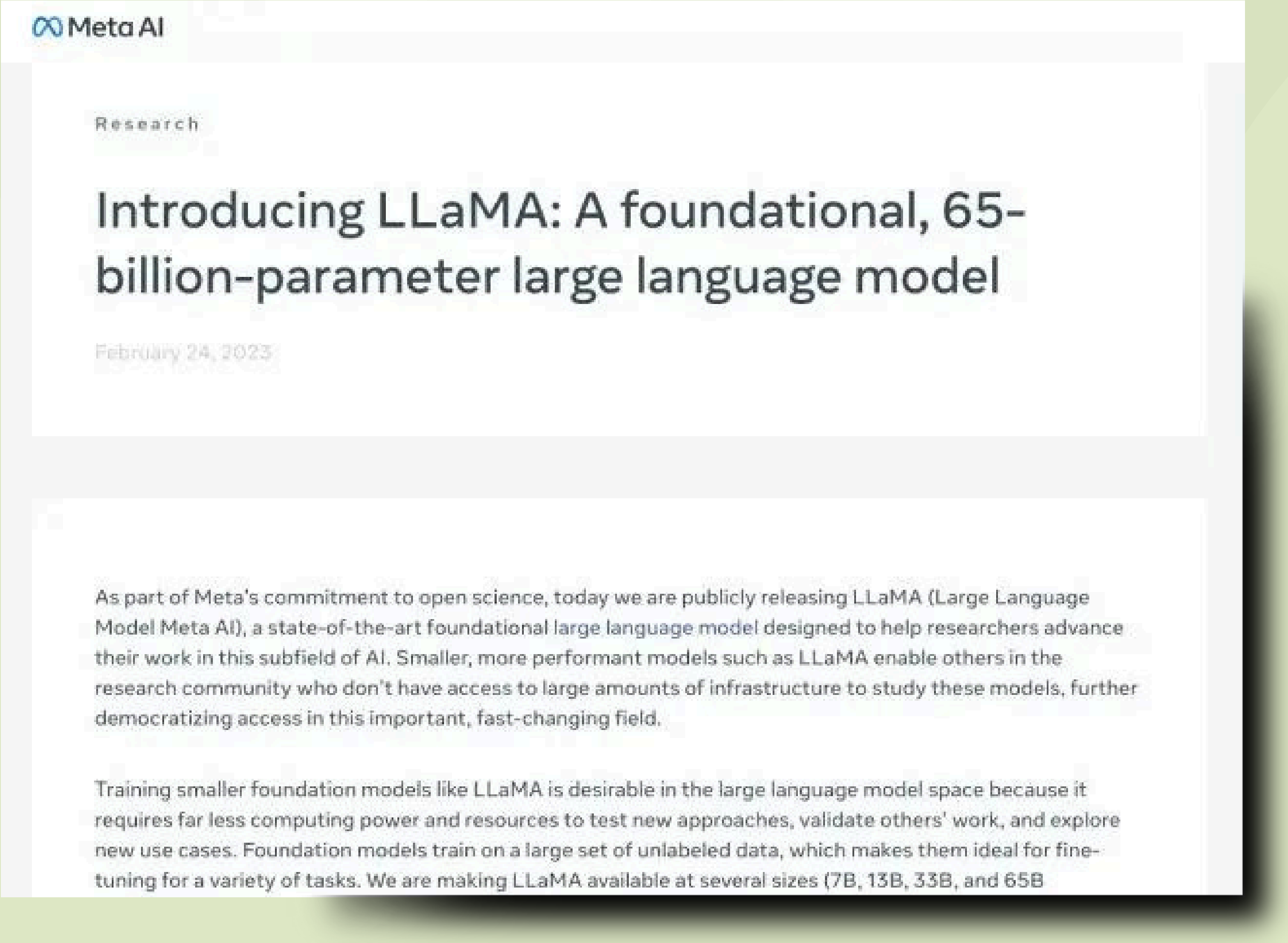
👀 **Learn more about Falcon LLM:** falconllm.tii.ae

Intended Use: this demo is intended to showcase an early finetuning of Falcon-40B, to illustrate the impact (and limitations) of finetuning on a dataset of conversations and instructions. We encourage the community to further build upon the base model, and to create even better instruct/chat versions!

⚠ **Limitations**: the model can and will produce factually incorrect information, hallucinating facts and actions. As it has not undergone any advanced tuning/alignment, it can produce problematic outputs, especially if prompted to do so. Finally, this demo is limited to a session length of about 1,000 words.

9. LLAMA

Ever since LLaMA models leaked online, Meta has gone all-in on open-source. It officially released LLaMA models in various sizes, from 7 billion parameters to 65 billion parameters. According to Meta, its LLaMA-13B model outperforms the GPT-3 model from OpenAI which has been trained on 175 billion parameters. Many developers are using LLaMA to fine-tune and create some of the best open-source models out there. Having said that, do keep in mind, LLaMA has been released for research only and can’t be used commercially unlike the Falcon model by the TII.



Meta AI

Research

Introducing LLaMA: A foundational, 65-billion-parameter large language model

February 24, 2023

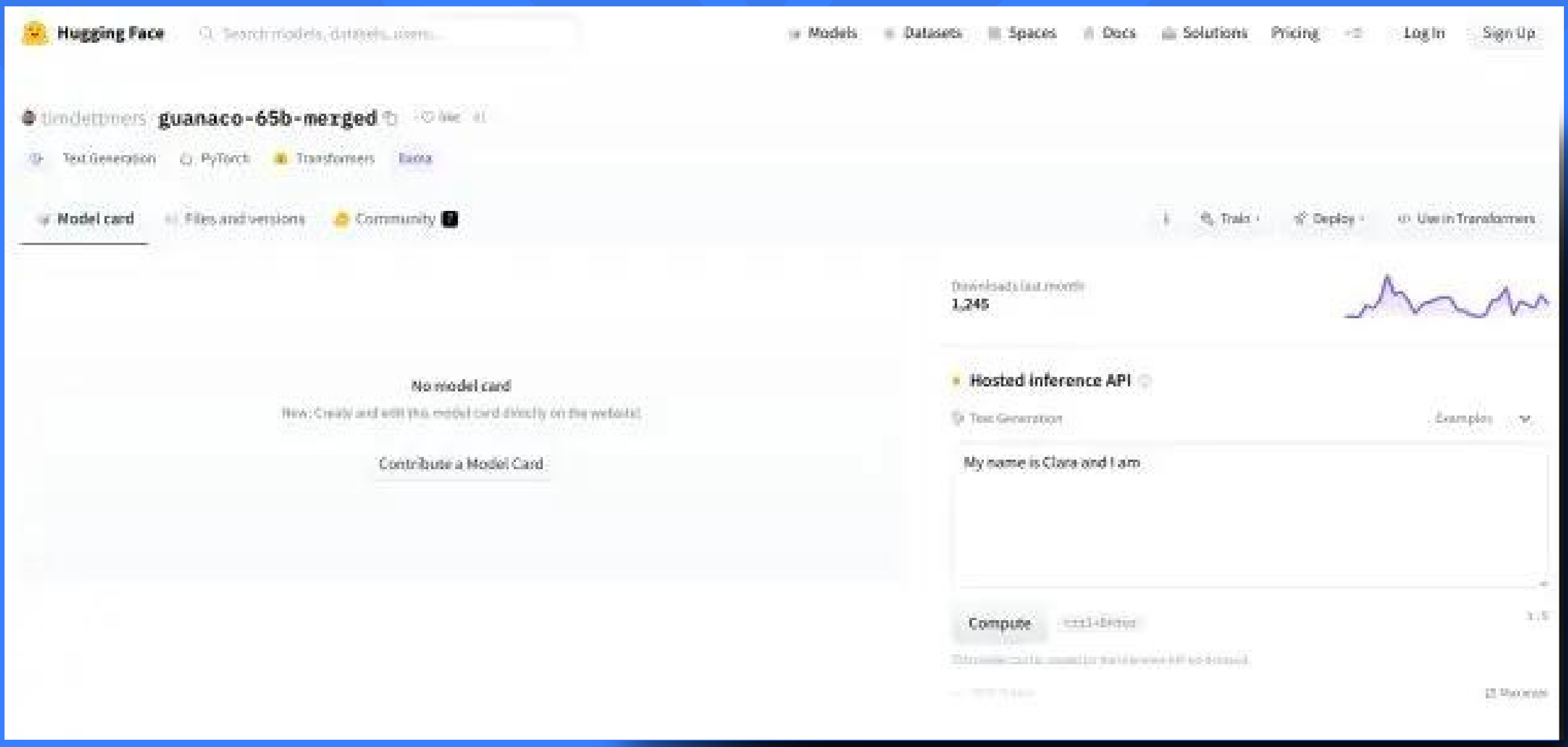
As part of Meta's commitment to open science, today we are publicly releasing LLaMA (Large Language Model Meta AI), a state-of-the-art foundational large language model designed to help researchers advance their work in this subfield of AI. Smaller, more performant models such as LLaMA enable others in the research community who don't have access to large amounts of infrastructure to study these models, further democratizing access in this important, fast-changing field.

Training smaller foundation models like LLaMA is desirable in the large language model space because it requires far less computing power and resources to test new approaches, validate others' work, and explore new use cases. Foundation models train on a large set of unlabeled data, which makes them ideal for fine-tuning for a variety of tasks. We are making LLaMA available at several sizes (7B, 13B, 33B, and 65B

PREV NEXT

10. GUANACO-65B

Among the several LLaMA-derived models, Guanaco-65B has turned out to be the best open-source LLM, just after the Falcon model. In the MMLU test, it scored 52.7 whereas the Falcon model scored 54.1. Similarly, in the TruthfulQA evaluation, Guanaco came up with a 51.3 score and Falcon was a notch higher at 52.5. There are four flavors of Guanaco: 7B, 13B, 33B, and 65B models. All of the models have been fine-tuned on the OASST1 dataset by Tim Dettmers and other researchers.



timdettmers / guanaco-65b-merged

Downloads last month
1,245

No model card

New: Create and edit this model card directly on the website!

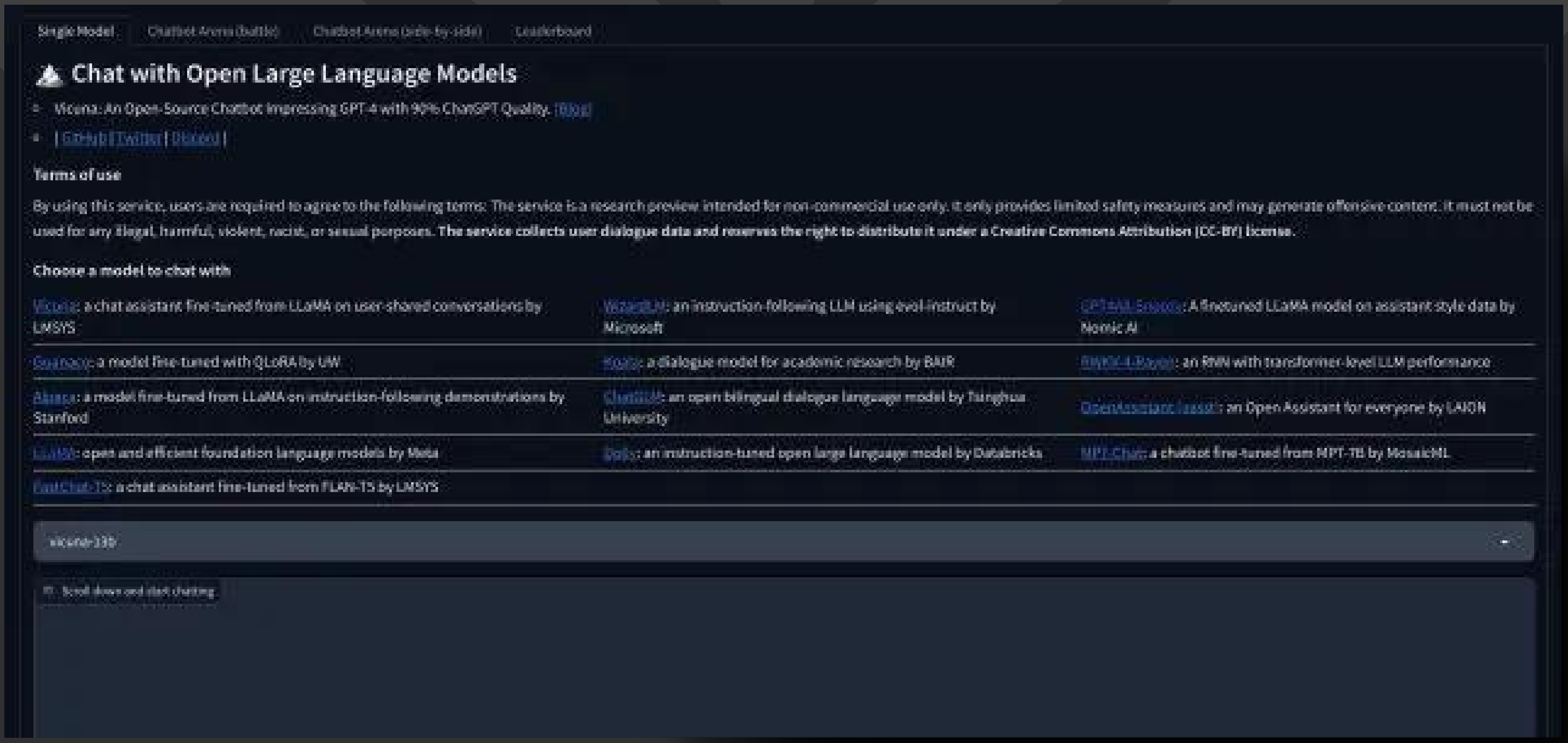
My name is Clara and I am

PREV

NEXT

11. VICUNA 33B

Vicuna is another powerful open-source LLM that has been developed by LMSYS. It has been derived from LLaMA like many other open-source models. It has been fine-tuned using supervised instruction and the training data has been collected from sharegpt.com, a portal where users share their incredible ChatGPT conversations. It’s an auto-regressive large language model and is trained on 33 billion parameters.



Chat with Open Large Language Models

- Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality. [Blog]
- | GitHub | Twitter | Discord |

Terms of use

By using this service, users are required to agree to the following terms: The service is a research preview intended for non-commercial use only. It only provides limited safety measures and may generate offensive content. It must not be used for any illegal, harmful, violent, racist, or sexual purposes. **The service collects user dialogue data and reserves the right to distribute it under a Creative Commons Attribution (CC-BY) license.**

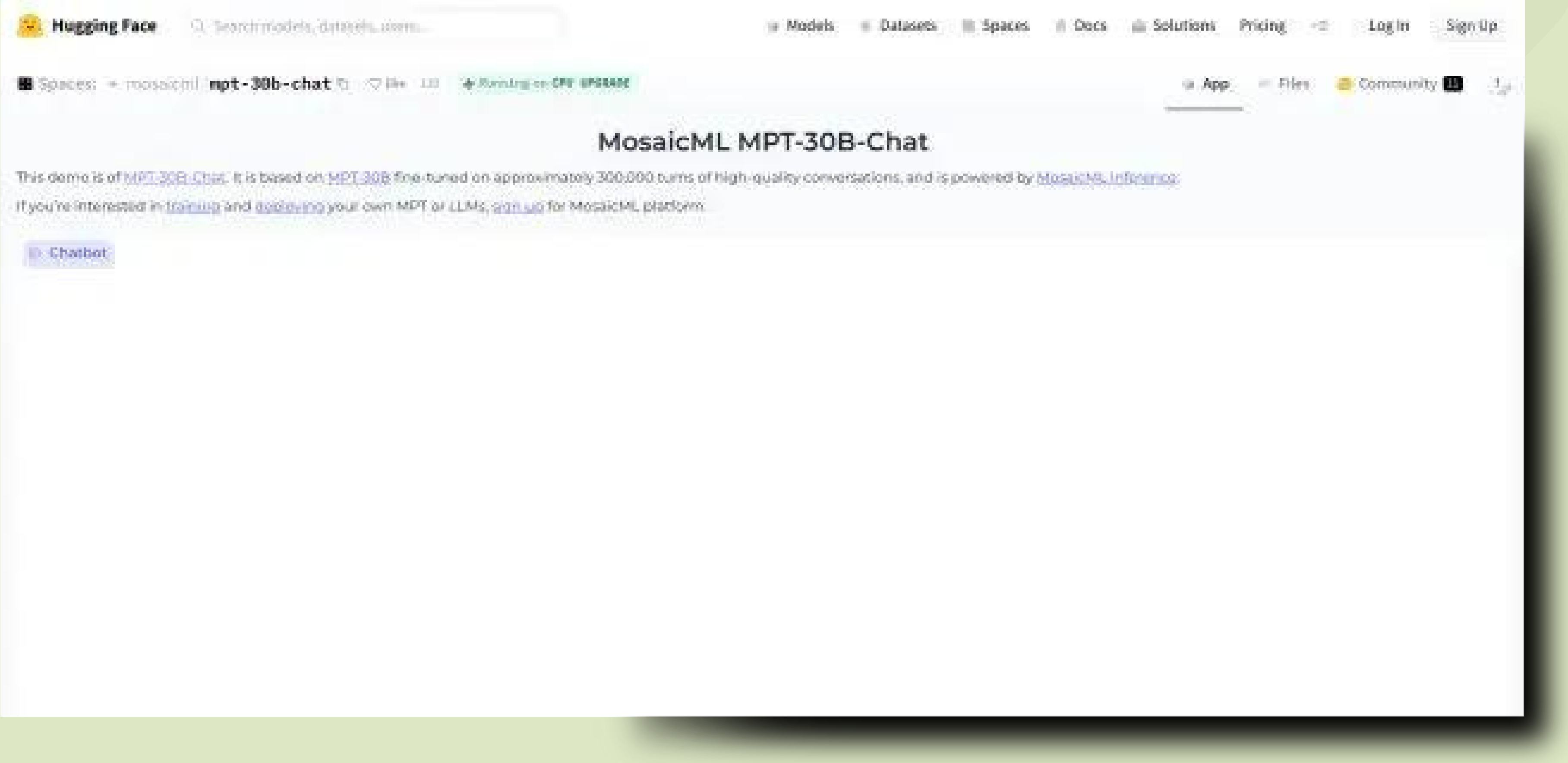
Choose a model to chat with

Vicuna: a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS	WizardLM: an instruction-following LLM using evol-instruct by Microsoft	GPT4All-Snoozy: A finetuned LLaMA model on assistant style data by Nomic AI
Guanaco: a model fine-tuned with QLoRA by UW	Koala: a dialogue model for academic research by BAIR	RWKV-4-Raven: an RNN with transformer-level LLM performance
Alpaca: a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford	ChatGLM: an open bilingual dialogue language model by Tsinghua University	OpenAssistant (oasst): an Open Assistant for everyone by LAION
LLaMA: open and efficient foundation language models by Meta	Dolly: an instruction-tuned open large language model by Databricks	MPT-Chat: a chatbot fine-tuned from MPT-7B by MosaicML
FastChat-T5: a chat assistant fine-tuned from FLAN-T5 by LMSYS		

vicuna-33b

12. MPT-30B

MPT-30B is another open-source LLM that competes against LLaMA-derived models. It has been developed by Mosaic ML and fine-tuned on a large corpus of data from different sources. It uses datasets from ShareGPT-Vicuna, Camel-AI, GPTeacher, Guanaco, Baize, and other sources. The best part about this open-source model is that it has a context length of 8K tokens.



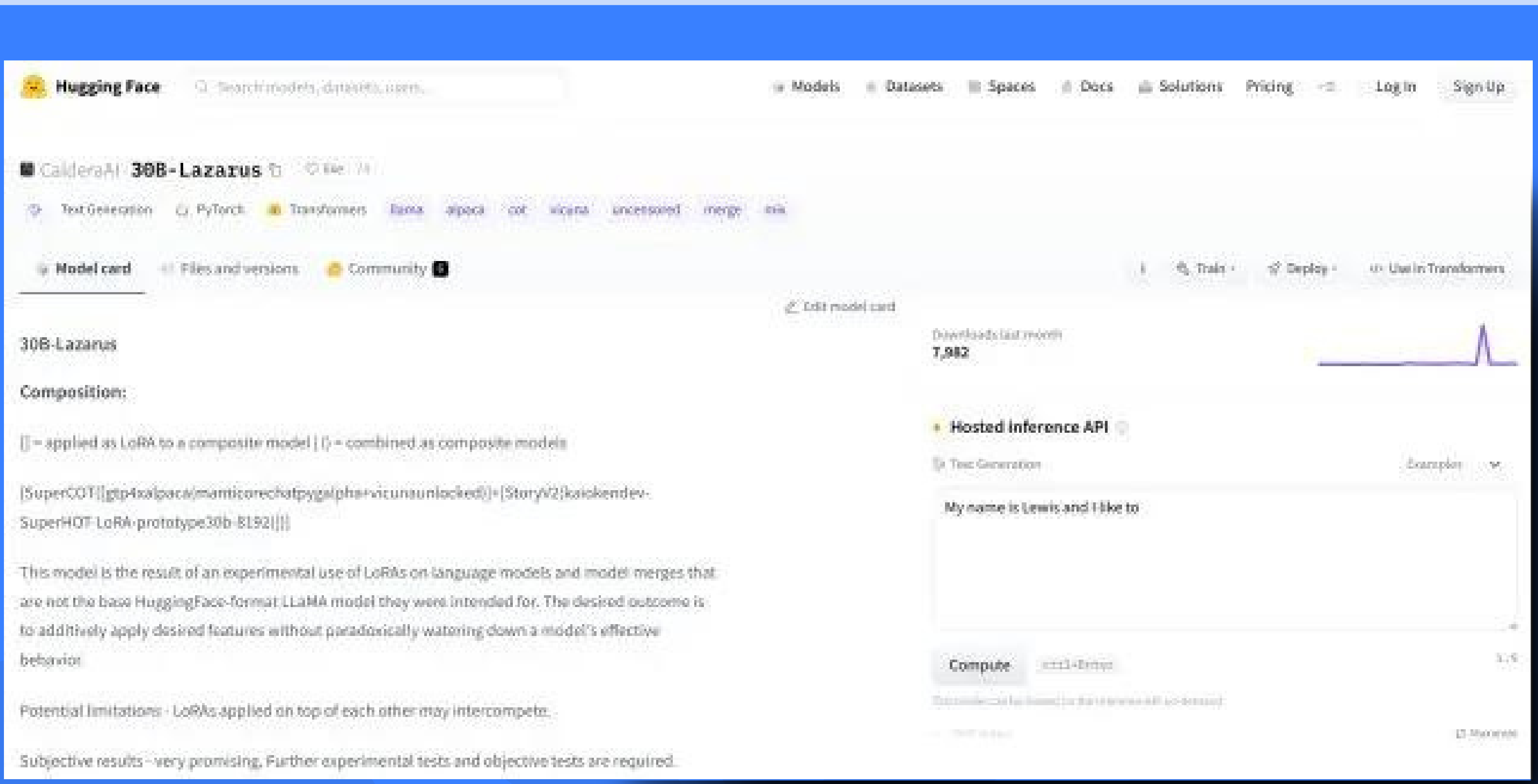
MosaicML MPT-30B-Chat

This demo is of MPT-30B-Chat. It is based on MPT-30B fine-tuned on approximately 300,000 turns of high-quality conversations, and is powered by MosaicML Inference.

If you're interested in training and deploying your own MPT or LLMs, sign up for MosaicML platform.

Dr. Martha Boeckenfeld

13. 30B-LAZARUS



30B-Lazarus

Composition:

[] = applied as LoRA to a composite model | () = combined as composite models

[SuperCOT([gtp4xalpaca(manticorechatpygalpha+vicunaunlocked)]+[StoryV2(kaiokendev-SuperHOT-LoRA-prototype30b-8192)])]

This model is the result of an experimental use of LoRAs on language models and model merges that are not the base HuggingFace-format LLaMA model they were intended for. The desired outcome is to additively apply desired features without paradoxically watering down a model's effective behavior.

Potential limitations - LoRAs applied on top of each other may intercompete.

Subjective results - very promising. Further experimental tests and objective tests are required.

My name is Lewis and I like to

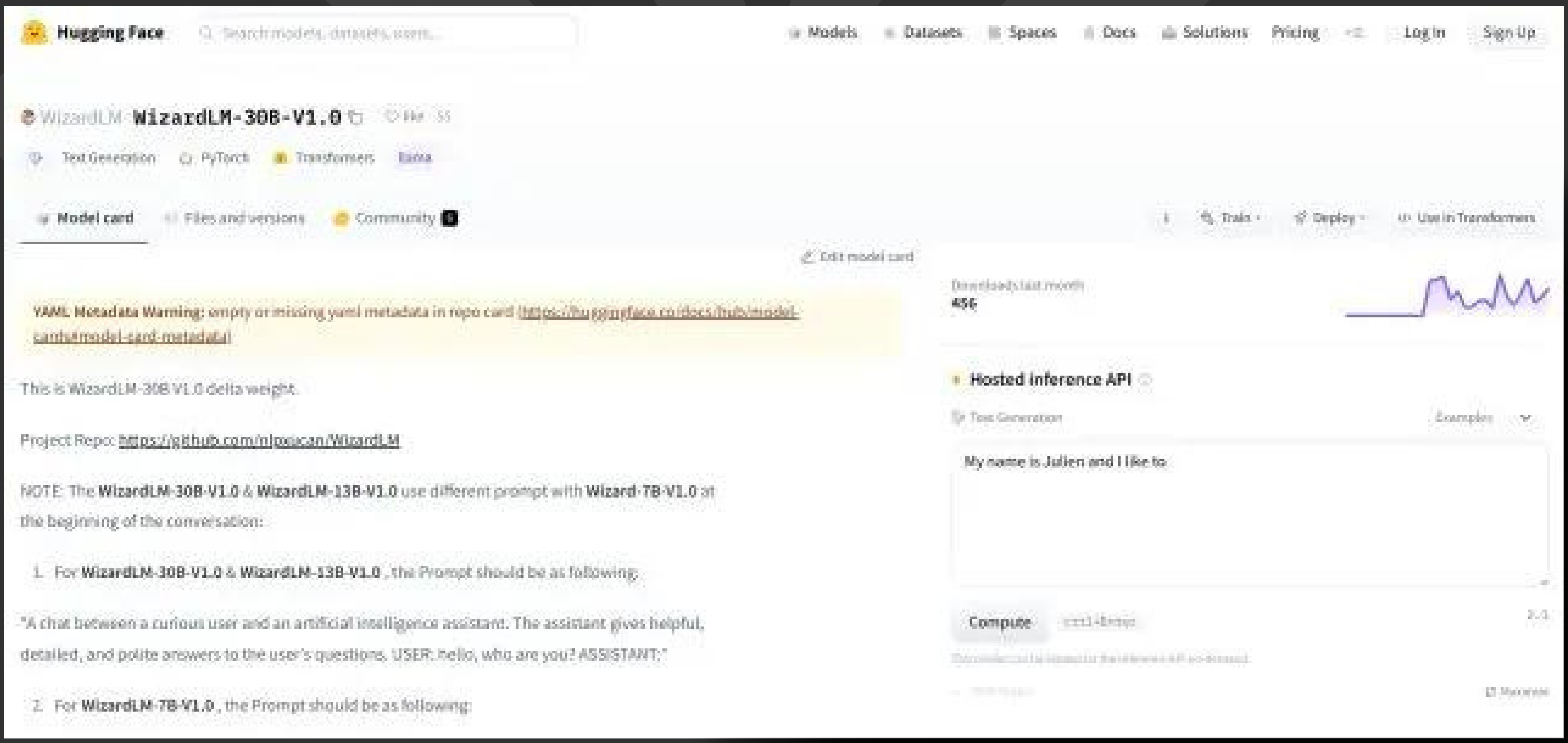
MPT-30B is another open-source LLM that competes against LLaMA-derived models. It has been developed by Mosaic ML and fine-tuned on a large corpus of data from different sources. It uses datasets from ShareGPT-Vicuna, Camel-AI, GPTeacher, Guanaco, Baize, and other sources. The best part about this open-source model is that it has a context length of 8K tokens.

PREV

NEXT

14. WIZARDLM

WizardLM is our next open-source large language model that is built to follow complex instructions. A team of AI researchers has come up with an Evol-instruct approach to rewrite the initial set of instructions into more complex instructions. And the generated instruction data is used to fine-tune the LLaMA model.



YAML Metadata Warning: empty or missing yaml metadata in repo card (https://huggingface.co/docs/hub/model-cards#model-card-metadata)

This is WizardLM-30B V1.0 delta weight.

Project Repo: https://github.com/nlpxucan/WizardLM

NOTE: The **WizardLM-30B-V1.0** & **WizardLM-13B-V1.0** use different prompt with **Wizard-7B-V1.0** at the beginning of the conversation:

1. For **WizardLM-30B-V1.0** & **WizardLM-13B-V1.0** , the Prompt should be as following:

"A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. USER: hello, who are you? ASSISTANT:"

2. For **WizardLM-7B-V1.0** , the Prompt should be as following:

My name is Julien and I like to

Dr. Martha Boeckenfeld

TOP 100 WOMEN OF THE FUTURE | WEB3 ADVISOR & INVESTOR | UN PEACE AMBASSADOR | FOUNDER MARTHAVERSE
- GUIDING EXECUTIVES & TEAMS FROM WEB2 TO WEB3 WITH STRATEGY, CREATION & EDUCATION

FOUND THIS POST HELPFUL ?

FOLLOW ME

FOR MORE INSIGHTFUL POSTS

CHECKOUT MORE ABOUT ME

https://linktr.ee/marthaverse

CLICK HERE