

# Large Language models

# what is a large language model

- Models like ChatGPT, Llama
- Predict the next token (word)

“Hello, my name is Eric and I ...”

“Hello, my name is Eric and I like”

# how are we going to do this

- We are using [Llama 3](#)
    - Kind of open source - model weights are public and you can do nearly anything with them
    - Rivals ChatGPT while being free, scoring 97% on gsm8k (vs 94%)
1. Get access to the model
  2. Code
  3. Profit

getting access to  
Llama 3

# creating an account

- HuggingFace contains many large language models
- Go to [huggingface.co](https://huggingface.co) and make a new account
  - School email is ok, you won't need to receive emails from them

# accessing the model

- Type meta-llama/Meta-Llama-3-8B-Instruct in the search box
- And click on the result

The screenshot shows the Hugging Face website interface. At the top left is the Hugging Face logo. A search bar contains the text "meta-llama/Meta-Llama-3-8B-Instruct". Below the search bar, a dropdown menu is open, displaying search results categorized into "Models", "Datasets", and "Spaces". The "Models" section is highlighted in blue, and the first result, "meta-llama/Meta-Llama-3-8B-Instruct", is selected. Other results include "mradermacher/meta-llama-Meta-Llama-3-8B-Instruct-fir" and "RichardErkhov/meta-llama\_-\_Meta-Llama-3-8B-Instruct". The "Datasets" section shows "open-llm-leaderboard/meta-llama\_\_Meta-Llama-3-8B-In..." and "llm-values/meta\_llama\_\_Meta\_Llama\_3\_8B\_Instruct\_ansv". The "Spaces" section shows "ImzanzamanML/meta-llama-Meta-Llama-3-8B-Instruct", "Funbi/meta-llama-Meta-Llama-3-8B-Instruct", and "carbuxer/meta-llama-Meta-Llama-3-8B-Instruct". A "Refresh List" button is visible next to the search results. On the right side of the page, there is a "Trending" section for the last 7 days, listing models like "stabilityai/stable-diffusion-3.5-large", "genmo/mochi-1-preview", "microsoft/OmniParser", and "nvidia/Llama-3.1-Nemotron-70B-Instruct...". The bottom right corner shows a notification for 489 likes.

# accessing the model

You should now be on the model's page, but you need to request access first.

**Hugging Face** Search models, datasets, users...

Models Datasets Spaces Posts Docs Solutions Pricing Log In Sign Up

meta-llama / **Meta-Llama-3-8B-Instruct** like 3.55k Follow Meta Llama 733

Text Generation Transformers Safetensors PyTorch English llama facebook meta llama-3 conversational text-generation-inference Inference Endpoints License: llama3

Model card Files and versions Community 190 Train Deploy Use this model

A newer version of this model is available: [meta-llama/Llama-3.1-8B-Instruct](#)

**You need to agree to share your contact information to access this model**

The information you provide will be collected, stored, processed and shared in accordance with the [Meta Privacy Policy](#).

**META LLAMA 3 COMMUNITY LICENSE AGREEMENT**

Meta Llama 3 Version Release Date: April 18, 2024

"Agreement" means the terms and conditions for use, reproduction, distribution and modification of the Llama Materials set forth herein.

"Documentation" means the specifications, manuals and documentation accompanying Meta Llama 3 distributed by Meta at <https://llama.meta.com/get-started/>.

"Licensee" or "you" means you, or your employer or any other person or entity (if you are entering into this Agreement on such person or...)

Log in or Sign Up to review the conditions and access this model content.

Downloads last month  
2,085,654

Safetensors Model size 8.03B params Tensor type BF16

**Inference API** Warm

Text Generation Examples

```
def fib(n):  
    if n <= 0:  
        return "Input should be a positive integer"  
    elif n == 1:  
        return 0  
    elif n == 2:  
        return 1  
    else:  
        return fibonacci(n-1) + fibonacci(n-2)
```

This function works by recursively calling itself with n-1 and n-2 until it reaches

# accessing the model

Fill out the request

- You don't need to use your real birthday, just pretend you're 18)
- Affiliation can be "student"


\* Reporting violations of the Acceptable Use Policy or unlicensed uses of Meta Llama 3: [LlamaUseReport@meta.com](mailto:LlamaUseReport@meta.com)

By agreeing you accept to share your contact information (email and username) with the repository authors.


**First Name**

**Last Name**

**Date of birth**

**Country**

**Affiliation**

Your country and region (based on approximate Internet address) will be shared with the model owner.

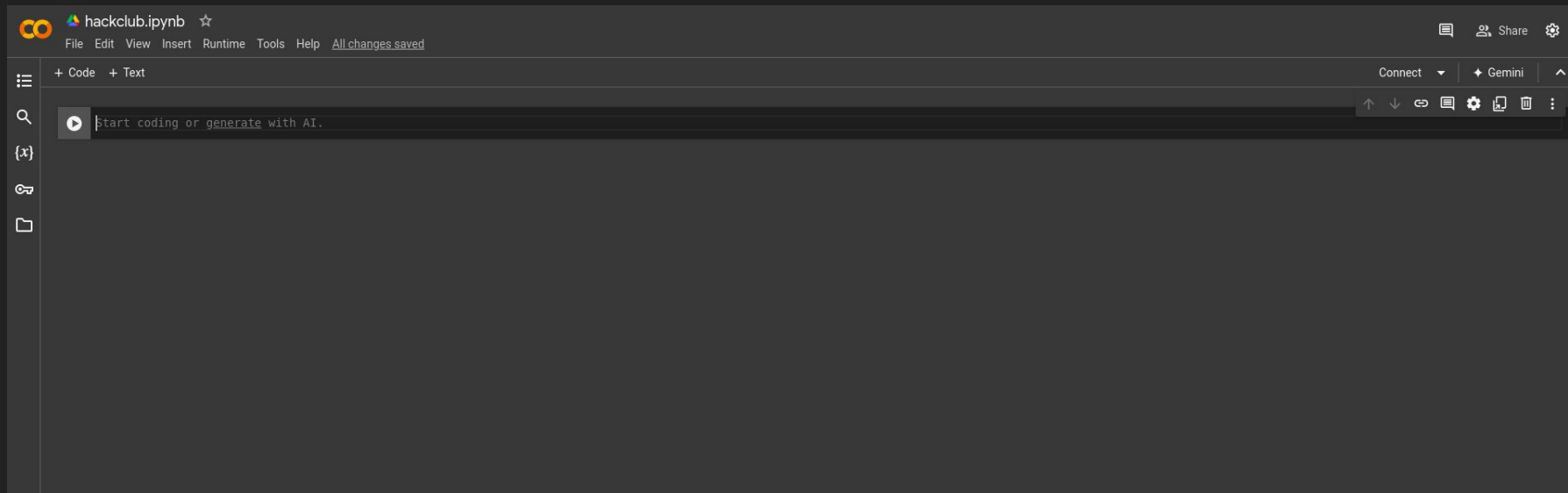


this will take a few hours to  
process, so in the meantime, let's  
get to coding

# opening a new google colab

Go to [colab.research.google.com](https://colab.research.google.com) and create a new project

Google Colab is an online code editor, and it also lets us use powerful GPUs to run our language model



## adding our setup code

In the box for code, add this:

```
import transformers
```

```
import torch
```

```
hf_token = ''
```

```
model_name = 'meta-llama/Meta-Llama-3-8B-Instruct'
```

## adding our setup code

Add another code box, and type

```
!pip install bitsandbytes
```

## adding our setup code

Add another code box, and add:

```
bnb_config = transformers.BitsAndBytesConfig(load_in_4bit=True,
bnb_4bit_use_double_quant=True, bnb_4bit_quant_type="nf4",
bnb_4bit_compute_dtype=torch.bfloat16)

model = transformers.AutoModelForCausalLM.from_pretrained(model_name,
device_map='auto', quantization_config=bnb_config, token=hf_token)

tokenizer = transformers.AutoTokenizer.from_pretrained(model_name,
token=hf_token)

generator = transformers.pipeline("text-generation", model=model,
tokenizer=tokenizer, pad_token_id=tokenizer.eos_token_id)
```

## add our main code

Add a fourth code box with this code:

```
chat_history = []  
  
while True:  
    next_user_input = input(' >')  
    chat_history.append({'role': 'user', 'content':  
next_user_input})  
    next_chat = generator(chat_history)[-1]['generated_text'][-1]  
    print(next_chat)  
    chat_history.append(next_chat)
```

# getting our HF api key

Go back to huggingface.co, click on your profile picture, and then settings

Then click on Access Tokens

The screenshot shows the Hugging Face user interface. At the top, there is a navigation bar with 'Docs', 'Solutions', and 'Pricing'. Below this is a yellow banner with the text 'Hugging Face is way more fun with friends and colleagues! 🥳 [Join an organization](#)' and a 'Dismiss this message' button. The main content area is divided into two sections. On the left is a user profile card for 'Eric' (greatericontop) with a menu of options: Profile, Account, Authentication, Organizations, Billing, Access Tokens (highlighted), and SSH and GPG Keys. On the right is the 'Access Tokens' page, which includes a '+ Create new token' button and a table of existing tokens. The table has columns for Name, Value, Last Refreshed Date, Last Used Date, and Permissions. One token is listed: 'llama2 token' with value 'hf\_...qmCL', last refreshed on 'Aug 3', and last used '4 minutes ago'. The permissions are 'FINEGRAINED'. A dropdown menu is open from the profile picture, showing options: Profile, greatericontop, Notifications, Inbox (0), New Model, New Dataset, New Space, New Collection, Create organization, Settings, Access Tokens, Billing, and Sign Out.

Hugging Face is way more fun with friends and colleagues! 🥳 [Join an organization](#) Dismiss this message

**Eric**  
greatericontop

Profile  
Account  
Authentication  
Organizations  
Billing  
**Access Tokens**  
SSH and GPG Keys

**Access Tokens**

User Access Tokens + Create new token

Access tokens authenticate your identity to the Hugging Face Hub and allow applications to perform actions based on token permissions.  
⚠️ Do not share your **Access Tokens** with anyone; we regularly check for leaked Access Tokens and remove them immediately.

Name	Value	Last Refreshed Date	Last Used Date	Permissions
llama2 token	hf_...qmCL	Aug 3	4 minutes ago	FINEGRAINED

Profile  
greatericontop  
Notifications  
Inbox (0)  
+ New Model  
+ New Dataset  
+ New Space  
+ New Collection  
Create organization  
Settings  
Access Tokens  
Billing  
Sign Out

## getting our HF api key

Create a new token, and activate the permission “Read access to contents of all public gated repos you can access”