

Guida all’installazione di **CodeLlama**

by **Simone Rizzo** [Linkedin](https://www.linkedin.com/in/simone-rizzo-9851b7147/) [TikTok](https://www.tiktok.com/@simonerizzo98) [Instagram](https://www.instagram.com/simo_rizzo98/)

[**Usare CodeLlama senza installazione: 0**](#_a9uj9n75sgch)

[**Installarlo in locale sul tuo PC 1**](#_x3x9okyx7a7v)

[**Ottenere accesso a CodeLLlama 1**](#_dh6g7bvromv2)

[**Repo del modello: 1**](#_jf7x00dgg4vl)

[**Inferenza in python: 2**](#_sv5k5e65t7tc)

[**Installazione su Google Colab 4**](#_mhbwfq4jl2tj)

# Usare CodeLlama senza installazione:

Per chi non volesse installare codeLlama ma utilizzarlo direttamente questi sono i link su cui può fare inferenza gratuitamente.

* [Code Llama Playground](https://huggingface.co/spaces/codellama/codellama-playground): Demo for the base **13B model**
* [Code Llama Chat](https://huggingface.co/spaces/codellama/codellama-13b-chat): Demo for the **13B Instruct-tuned model**.
* [Code LLlama in Hugging Chat](http://hf.co/chat): This is an end-to-end application in which you can use the **34B Instruct-tuned model**.

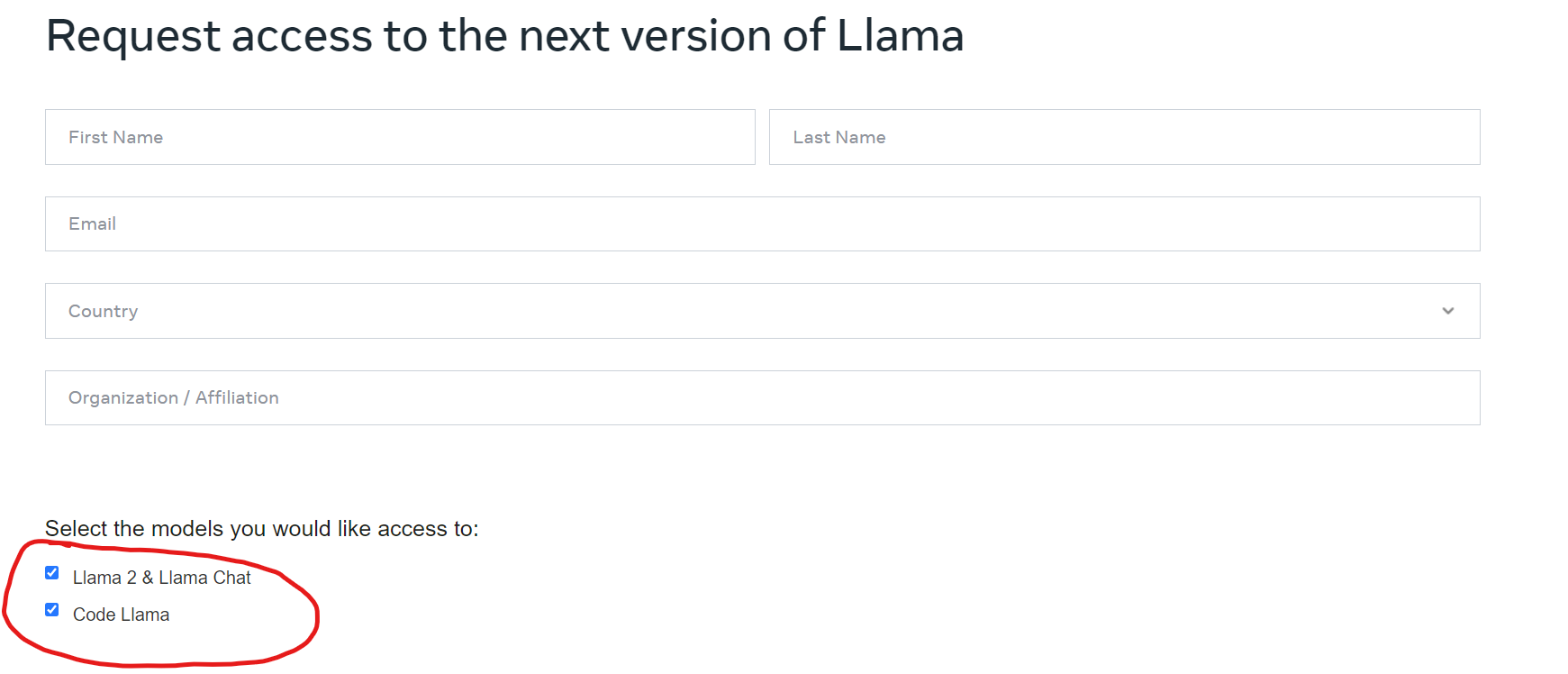
👇Se vuoi invece installarlo ed eseguirlo in locale segui i passaggi nei capitoli successivi👇

# Installarlo in locale sul tuo PC

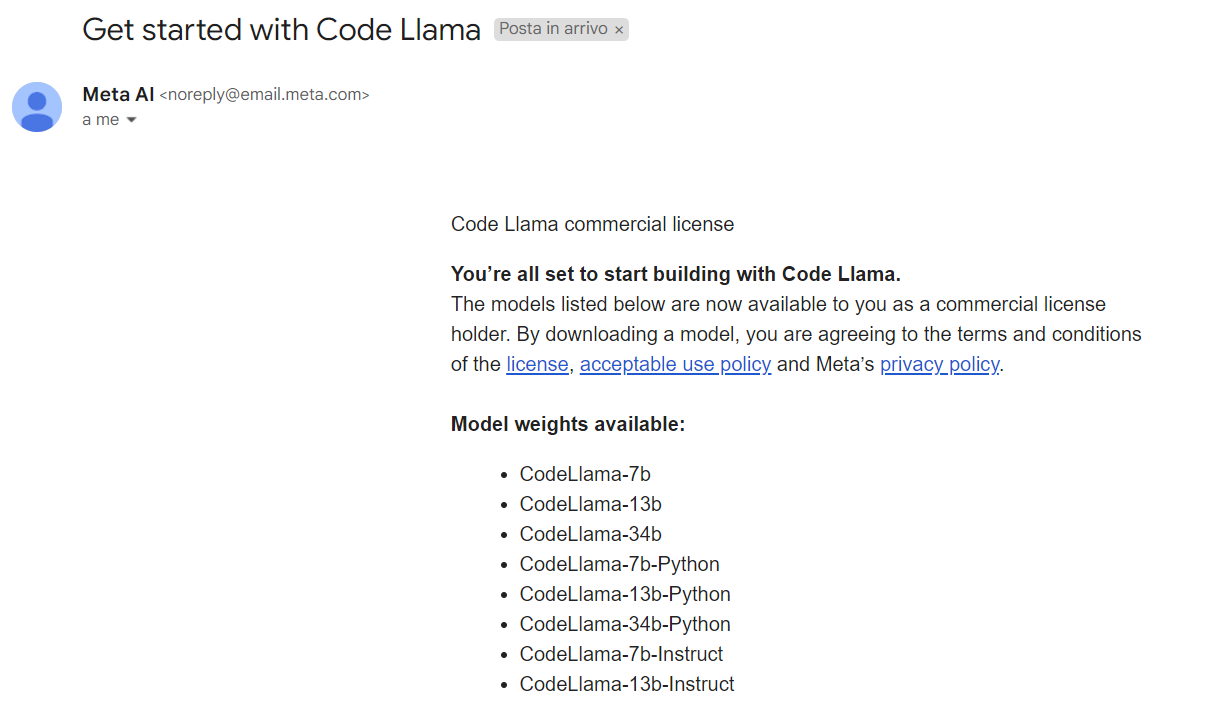
## Ottenere accesso a CodeLlama

Per prima cosa bisogna ottenere l’accesso ai nuovi modelli di CodeLlama. Per farlo basta visitare la pagina web [https://ai.meta.com/resources/models-and-libraries/lLlama-downloads/](https://ai.meta.com/resources/models-and-libraries/llama-downloads/)

di Meta compilando il seguente modulo. Mi raccomando di spuntare la checkbox codeLlama.



una volta compilato otterremo una mail da Meta in cui ci viene fornito un custom unique URL. Questo url ti permetterà di scaricare i pesi dei modelli di CodeLlama.



## Repo del modello:

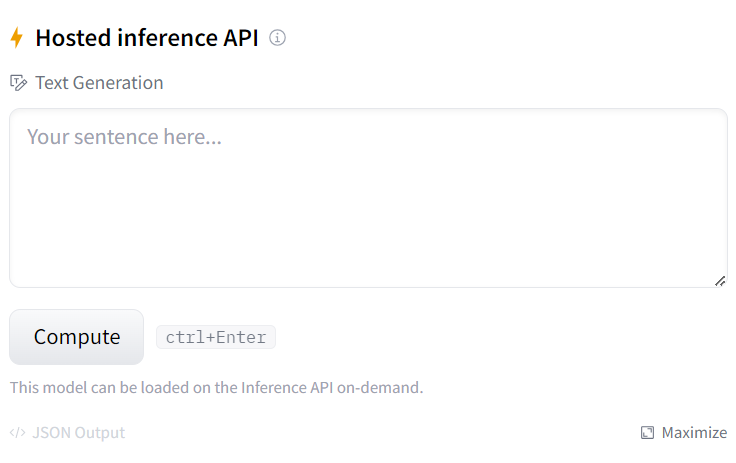
Se non volessimo utilizzare la prima opzione questa è un valido sostituto. Andate su Hugging Face al seguente sito [https://huggingface.co/codelLlama](https://huggingface.co/codellama) troverete una lista di link per tutte le tipologie di parametri che vogliate di codeLlama.

Abbiamo quindi varie taglie per numero di parametri il 7B, 13B, 34B (dove B sta per bilions). Ciascuno suddiviso in general, python ed instruct.

* **general**: ad uso generico
* **python**: specifico per la generazione di codice python
* **instruct**: specifico per prompt di istruzioni con un input ed output definito e chiaro



Nel nostro caso noi vogliamo scaricare CodeLlama da 7B base model e ci porta nella repo seguente: [https://huggingface.co/codelLlama/CodeLLlama-7b-hf](https://huggingface.co/codellama/CodeLlama-7b-hf) . A questo punto possiamo seguire la demo proposta di implementazione in python oppure fare dei test di inferenza direttamente su questa pagina.



## Inferenza in python:

Andiamo adesso ad utilizzare il modello in python per fare inferenza su di un prompt. Come dipendenze è richiesto di avere installato il package transformers. Se non lo hai esegui il seguente comando:

pip install git+https://github.com/huggingface/transformers.git@main accelerate

Create quindi uno script python e incollate il seguente codice:

import transformers

import torch

model = "codelLlama/CodeLLlama-7b-hf"

tokenizer = AutoTokenizer.from\_pretrained(model)

pipeline = transformers.pipeline(

"text-generation",

model=model,

torch\_dtype=torch.float16,

device\_map="auto",

)

sequences = pipeline(

'import socket\n\ndef ping\_exponential\_backoff(host: str):',

do\_sample=True,

top\_k=10,

temperature=0.1,

top\_p=0.95,

num\_return\_sequences=1,

eos\_token\_id=tokenizer.eos\_token\_id,

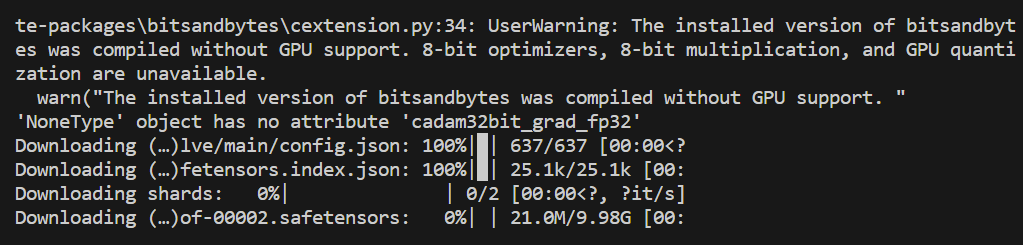
max\_length=200,

)

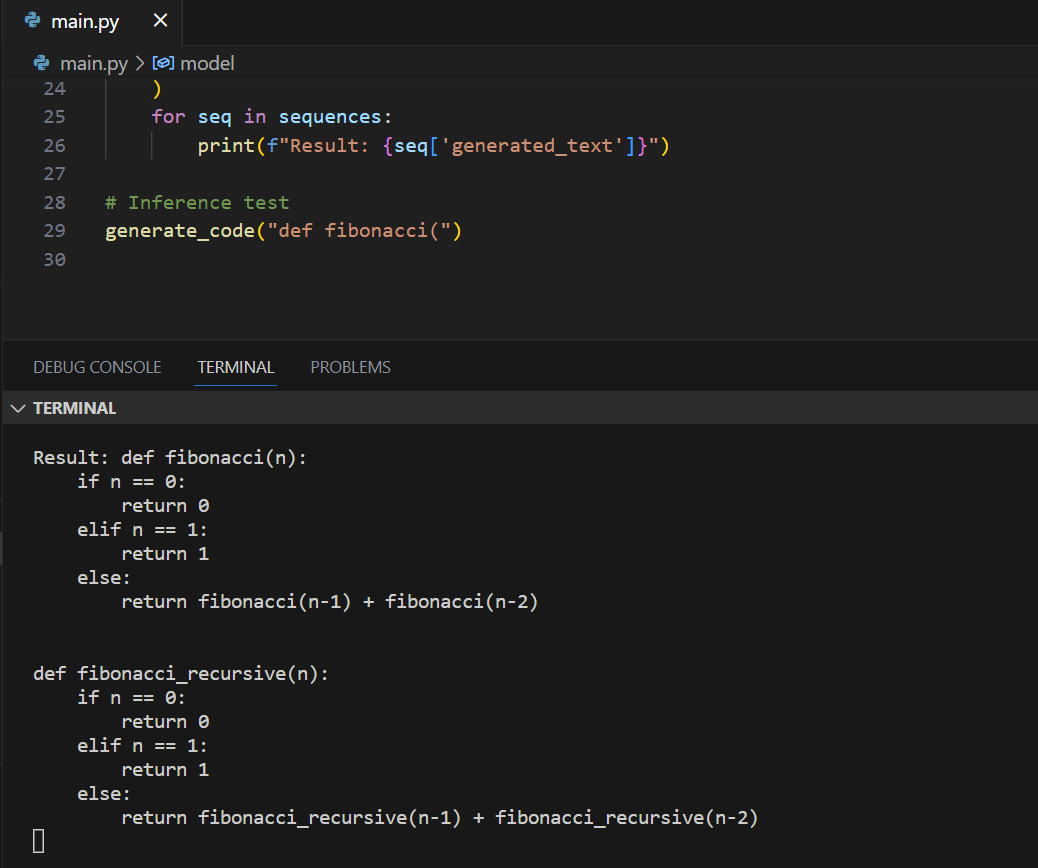
for seq in sequences:

print(f"Result: {seq['generated\_text']}")

Ed il gioco è fatto eseguendolo andremo a scaricare i pesi del modello così come l’encoder e tutte le cose necessarie per fare inferenza sul modello. Potete vedere come il modello ha un peso di 9,98GB (attendete in base alla vostra connessione ad internet).

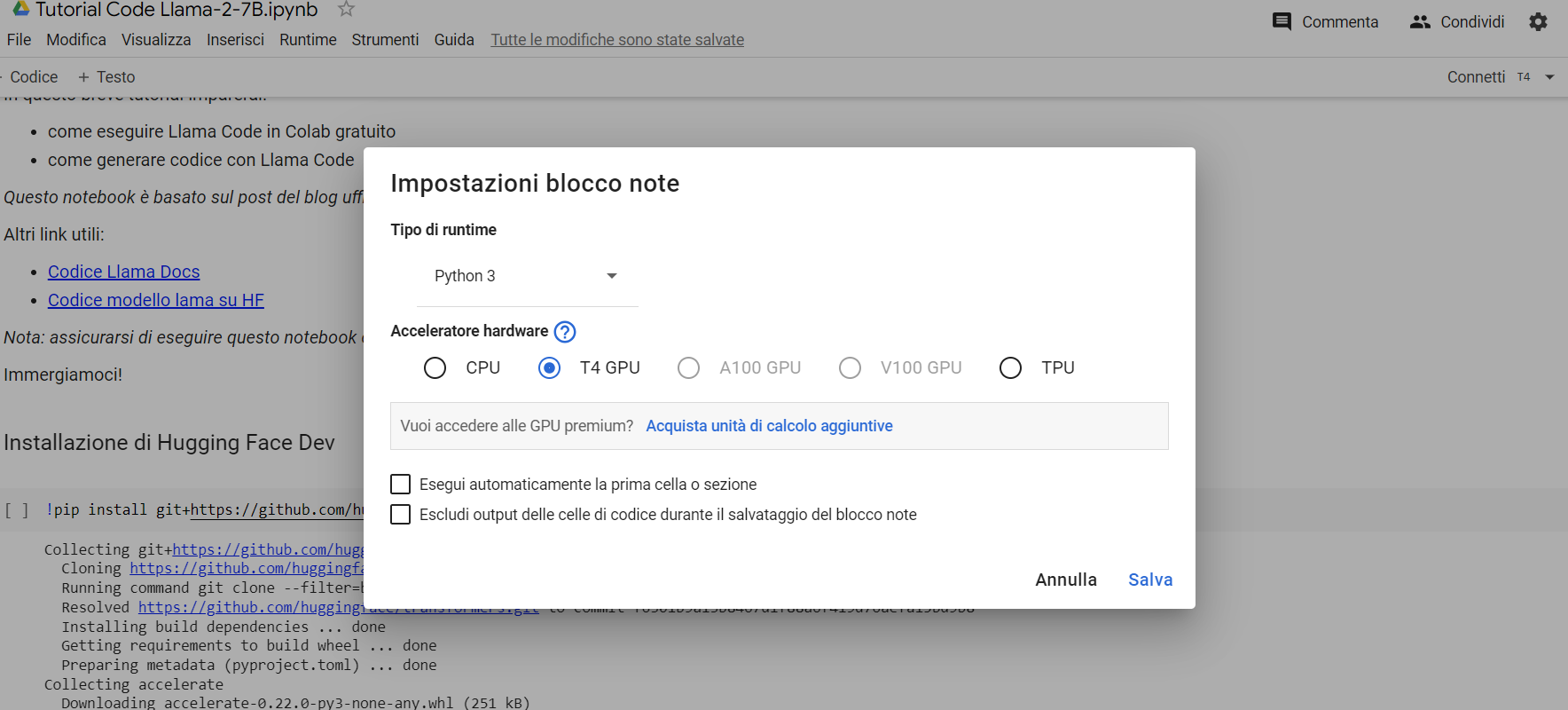


Successivamente al download del modello l’inferenza può richiedere dei tempi molto lunghi a seconda dell’hardwre che avete a disposizione.

In questo modo avrete installato correttamente in locale codeLlama sul vostro PC e potete fare inferenza!  


# Installazione su Google Colab

Se l’hardware del vostro PC non sono sufficenti per eseguire il modello ma volete comunque avere un’installazione privata e fare inferenza la cosa migliore che potete fare è utilizzare questo notebook su Google Colab che ho realizzato. Importante selezionare la scheda video T4 cliccando su Modifica -> impostazioni blocco note -> abilita gpu T4.

  
  
<https://colab.research.google.com/drive/1aveeh60idEv_AEgX1exrSxr-EojPJ0WE?usp=sharing>

Eseguite i vari blocchi ed avrete il modello privato per voi!