

```
!nvidia-smi
```

```
Mon Sep 22 17:38:53 2025
```

| NVIDIA-SMI 550.54.15 | | | | Driver Version: 550.54.15 | | CUDA Version: 12.4 | |
|----------------------|----------|---------------|------------------|---------------------------|----------|--------------------|-----|
| GPU | Name | Persistence-M | Bus-Id | Disp.A | Volatile | Uncorr. ECC | |
| Fan | Temp | Perf | Pwr:Usage/Cap | Memory-Usage | GPU-Util | Compute M. MIG M. | |
| 0 | Tesla T4 | Off | 00000000:00:04.0 | Off | 0 | | |
| N/A | 62C | P8 | 10W / 70W | 0MiB / 15360MiB | 0% | Default | N/A |

| Processes: | GPU | GI | CI | PID | Type | Process name | GPU Memory Usage |
|----------------------------|-----|----|----|-----|------|--------------|------------------|
| No running processes found | | | | | | | |

```
!pip install transformers
```

```
Requirement already satisfied: transformers in /usr/local/lib/python3.12/dist-packages (4.56.1)
Requirement already satisfied: filelock in /usr/local/lib/python3.12/dist-packages (from transformers) (3.19.1)
Requirement already satisfied: huggingface-hub<1.0,>=0.34.0 in /usr/local/lib/python3.12/dist-packages (from transformers) (0.34.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.12/dist-packages (from transformers) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.12/dist-packages (from transformers) (25.0)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.12/dist-packages (from transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.12/dist-packages (from transformers) (2024.11.6)
Requirement already satisfied: requests in /usr/local/lib/python3.12/dist-packages (from transformers) (2.32.4)
Requirement already satisfied: tokenizers<0.23.0,>=0.22.0 in /usr/local/lib/python3.12/dist-packages (from transformers) (0.22.0)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.12/dist-packages (from transformers) (0.6.2)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.12/dist-packages (from transformers) (4.67.1)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.12/dist-packages (from huggingface-hub<1.0,>=0.34.0->transformers) (2025.9.2)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.12/dist-packages (from huggingface-hub<1.0,>=0.34.0->transformers) (4.12.2)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.3 in /usr/local/lib/python3.12/dist-packages (from huggingface-hub<1.0,>=0.34.0->transformers) (1.1.7)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.12/dist-packages (from requests->transformers) (3.4.0)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.12/dist-packages (from requests->transformers) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.12/dist-packages (from requests->transformers) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.12/dist-packages (from requests->transformers) (2025.11.12)
```

```
from transformers import AutoTokenizer
```

```
# Load the tokenizer for a specific model (e.g., GPT-2)
tokenizer = AutoTokenizer.from_pretrained("gpt2")
```

```
# Tokenize some input text
text = "Hello, how are you?"
tokens = tokenizer(text, return_tensors='pt')
print(tokens)
```

```
/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it as the secret `HF_TOKEN` in your Colab secrets, and restart this notebook.
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
warnings.warn(
{'input_ids': tensor([[15496, 11, 703, 389, 345, 30]]), 'attention_mask': tensor([[1, 1, 1, 1, 1, 1]])}
```

```
from transformers import AutoModelForCausalLM
```

```
# Load the pre-trained GPT-2 model
model = AutoModelForCausalLM.from_pretrained("gpt2")

# Generate text
input_ids = tokenizer.encode("indian cricket", return_tensors='pt')
output = model.generate(input_ids, max_length=50)
generated_text = tokenizer.decode(output[0], skip_special_tokens=True)

print(generated_text)
```

```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input as a pair of (input_ids, attention_mask) where attention_mask is of dtype torch.bool and of length input_ids. Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
The attention mask is not set and cannot be inferred from input because pad token is same as eos token. As a consequence, you may observe unexpected behavior. Please pass your input as a pair of (input_ids, attention_mask) where attention_mask is of dtype torch.bool and of length input_ids.
```

```
The team's captain, Ravi Shankar, has been in the country for over a decade.
```

```
The team's captain, Ravi Shankar,
```

[illegible]

```
from transformers import pipeline, set_seed
generator = pipeline('text-generation', model='gpt2')
set_seed(42)
generator("Hello, I'm a language model.", max length=30, num return sequences=5)
```

2/5

I mean, what are you doing?"\n\nA few words from the man, who did not return calls.\n\nHe said she was "overworked" and that he had "made a mistake." "I think there's not one right answer," he said, adding that he had been told there were "no more questions." \n\nThe writer said the problem stemmed from her job as a writer in the online publication The New Yorker, where she was a part-time writer and editor.\n\nShe said she had been told that while she was happy to work at The New Yorker, her "job at this time was to write fiction and I was not. I thought I could have a full-time job at The New Yorker. I was wrong." \n\nThe writer's employer did not respond to a request for comment Friday.\n\nThe writer was a co-founder of G.I. Joe's magazine and a co-founder of the online publishing company Ado, which has its own website.\n\nA copy of G.I. Joe's website listed her as "a contributing editor and contributing editor to [The New Yorker].",

```
{'generated_text': 'Hello, I\'m a language model, and it\'s not about me. It\'s about people.\n\nIf you\'re a person and you want to tell people what a language is, you have to be able to tell them what the language is about.\n\nLang, who came to the UK with her mother, has been studying English since she was 8.\n\nShe says she is passionate about how to understand people and how they use language.\n\nI\'m a language model, and it\'s not about me. It\'s about people. It\'s about people as far as I\'m concerned.\n\nShe says she\'s always been interested in learning English and how to express herself and the world around her.\n\nBut she also says she doesn\'t understand why some people don\'t understand her and her language.\n\nWhat do you get when you talk about the world of your language?\n\nYou get to know people and you know people speak more than you do, but you\'re not allowed to do that.\n\nSo you\'re not allowed to do that.\n\nTheresa May has repeatedly claimed she wants to "unite the world" and is working to create an "open-ended" international language system.\n\nBut the Government has'},
```

```
{'generated_text': 'Hello, I\'m a language model, not a language model. I\'m thinking of the languages in which we have formal semantics. One of my favorite languages is C#, which is the language of the language model. We\'re not talking about the semantics of a language model in a formal sense. We\'re talking about language models in which the language model is the only set of semantics that you can apply to any particular language.\n\nOne of the things I like about this kind of formal semantics is that it\'s a good way to develop a language model without having to go through languages that are not formal models. And I think you can do it with C#, which is not formal models.\n\nA lot of the things you will be interested in coming out of this are examples of non-formal semantics. I would like to talk about the second way that you can say, "I want to write this language model in C#.\n\nThere\'s a lot of things that you will be interested in. First of all, we have the language model. It\'s a language model, it\'s not a syntax model. We have a language model that we can do what we want. It\'s a language model that we can look at. It\'s a language model that you can apply to'}}]
```

```
from transformers import pipeline, set_seed
generator = pipeline('text-generation', model='gpt2')
set_seed(42)
generator("Hello, explain about indian economy ", max_length=10, num_return_sequences=2)
```

Device set to use cuda:0

Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True` to explicit Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.

Both `max_new_tokens` (=256) and `max_length` (=10) seem to have been set. `max_new_tokens` will take precedence. Please refer to [{}generated_text': 'Hello, explain about indian economy , for you understand that the Indian economy is highly dependent on the Indian Government. The Indian Government provides for its citizens with a comprehensive income tax, and this income tax is not charged in India, but it is charged on a monthly basis by the Indians. The Indian Government also charges tax on the income of foreign citizens. In other words, Indian citizens can do their own tax, but they could not do the same as Indian citizens in the United States where they are subject to the taxes imposed by the Indian Government. This is also why Indians are not allowed to move around freely.\n\nIn the United States, your Indian citizenship is not subject to taxation, but you own the land and the people of the United States.\n\nThe Indians do not have to pay any income tax on their own, they can move freely throughout the country.\n\n, they can move freely throughout the country. In other words, the Indian citizens can enjoy the benefits of the Indian economy, as long as they live in the same locality or live in the same land.\n\n, as long as they live in the same locality or live in the same land. If you live in the same locality with a foreigner, you can become a citizen of India.\n\n. If you live in'],

```
{'generated_text': "Hello, explain about indian economy , my blog post on this topic.\n\nI\'m a big fan of the game, but I guess it\'s hard to figure out why a person would spend their time and energy on it.\n\nI also don\'t like to think about it too much. I like to think about it as a concept that\'s really important to all people, not just my own. I think that it\'s important to me to be able to think about it as a concept that gets to a point where I feel comfortable with what I\'ve said about it.\n\nIf I had to guess, it would be that I\'d like to say that I think of indians as a country that\'s very rich, but if you ask me where I get my money, I don\'t think of it that way. I think of it as country where I have a lot of money. I do have money because I have a lot of friends, but I also have a lot of friends who like doing work and doing things that I do, as well as a lot of people who are very interested in the game.\n\nI think that it\'s very important for me to be able to say, in a sense, that I\'m a person who gets a lot of money, and that"]}]
```

```
from transformers import GPT2Tokenizer, GPT2Model
tokenizer = GPT2Tokenizer.from_pretrained('gpt2')
model = GPT2Model.from_pretrained('gpt2')
text = "Replace me by any text you'd like."
encoded_input = tokenizer(text, return_tensors='pt')
output = model(**encoded_input)
print(output)
```

```
BaseModelOutputWithPastAndCrossAttentions(last_hidden_state=tensor([[[[ 0.1629, -0.2166, -0.1410, ..., -0.2619, -0.0819, 0.005
[ 0.4628, 0.0248, -0.0785, ..., -0.0859, 0.5122, -0.3939],
[-0.0644, 0.1551, -0.6306, ..., 0.2488, 0.3691, 0.0833],
...,
[-0.5591, -0.4490, -1.4540, ..., 0.1650, -0.1302, -0.3740],
[ 0.1400, ..., -0.3875, -0.7916, ..., -0.1780, 0.1824, 0.2185],
[ 0.1721, -0.2420, -0.1124, ..., -0.1068, 0.1205, -0.3213]]]],
grad_fn=<ViewBackward0>), past_key_values=DynamicCache(layers=[DynamicLayer, DynamicLayer, DynamicLayer, DynamicLayer, I
```

```
from transformers import GPT2Tokenizer, TFPGPT2Model

tokenizer = GPT2Tokenizer.from_pretrained("gpt2")
model = TFPGPT2Model.from_pretrained("gpt2", from_pt=True)

text = "Replace me by any text you'd like."
encoded_input = tokenizer(text, return_tensors="tf")
```

```
output = model(encoded_input)
print(output)
```

pytorch_model.bin: 100% 548M/548M [00:10<00:00, 67.4MB/s]

All PyTorch model weights were used when initializing TFGPT2Model.

All the weights of TFGPT2Model were initialized from the PyTorch model.

If your task is similar to the task the model of the checkpoint was trained on, you can already use TFGPT2Model for prediction. TensorFlow and JAX classes are deprecated and will be removed in Transformers v5. We recommend migrating to PyTorch classes or TFBaseModelOutputWithPastAndCrossAttentions(last_hidden_state=<tf.Tensor: shape=(1, 10, 768), dtype=float32, numpy=

```
array([[[[ 0.16290577, -0.21657419, -0.14102745, ..., -0.26188618,
          -0.08190881,  0.00923978],
         [ 0.46279675,  0.02483805, -0.0785372 , ..., -0.08585826,
          0.51222366, -0.39390466],
         [-0.06436783,  0.15511821, -0.6305839 , ...,  0.24878348,
          0.36905405,  0.08326927],
         ...,
         [-0.5590812 , -0.44902402, -1.4539894 , ...,  0.16499005,
          -0.13022885, -0.37402722],
         [ 0.14001575, -0.38752818, -0.7915612 , ..., -0.1779689 ,
          0.18236114,  0.21849152],
         [ 0.17207077, -0.24204722, -0.11238727, ..., -0.1068422 ,
          0.1205473 , -0.3212943 ]]], dtype=float32)>), past_key_values=(<tf.Tensor: shape=(2, 1, 12, 10, 64), dtype=float32,
array([[[[[-1.07186723e+00,  2.41698909e+00,  9.66034472e-01, ...,
          -4.78705823e-01, -3.31556976e-01,  1.79252315e+00],
         [-2.28969359e+00,  2.54237032e+00,  8.31743181e-01, ...,
          -5.29929280e-01, -2.48278284e+00,  1.35369802e+00],
         [-2.28563309e+00,  2.71245575e+00,  2.47251630e+00, ...,
          -1.49111617e+00, -1.84272110e+00,  1.64932966e+00],
         ...,
         [-3.32033706e+00,  2.33251595e+00,  2.70611525e+00, ...,
          -1.15692115e+00, -1.55860960e+00,  2.40759659e+00],
         [-2.99168444e+00,  2.27005696e+00,  2.17415881e+00, ...,
          -8.66989911e-01, -1.64104259e+00,  1.92371106e+00],
         [-2.50655484e+00,  2.61395168e+00,  2.13473463e+00, ...,
          -6.27139658e-02, -2.05424690e+00,  1.65676892e+00]],
         [[ 4.79599476e-01, -1.13052562e-01, -1.48538733e+00, ...,
          1.16065383e+00,  1.84115338e+00,  1.36818612e+00],
         [-7.27268338e-01, -1.13618815e+00, -1.08502722e+00, ...,
          -6.73586965e-01,  3.26177359e+00,  2.09914953e-01],
         [-1.44409227e+00, -3.06468940e+00, -4.16119671e+00, ...,
          -1.47875154e+00,  3.27183390e+00, -2.80251622e-01],
         ...,
         [ 8.51459503e-01, -1.59904584e-01,  1.15671441e-01, ...,
          -8.95856857e-01,  4.11782742e+00,  7.13325799e-01],
         [-7.69265294e-02, -1.76725709e+00, -1.12068915e+00, ...,
          -1.62757242e+00,  3.10953665e+00,  1.02366257e+00],
         [-9.11841750e-01, -3.26741517e-01, -2.04092026e+00, ...,
          -3.52716893e-01,  1.16257918e+00,  3.73330027e-01]],
         [[-2.33778954e-01, -8.68849277e-01,  1.65420091e+00, ...,
          -1.59637439e+00, -1.56355262e+00,  1.09308207e+00],
         [ 3.69760990e-01,  4.92904007e-01,  1.41548944e+00, ...,
          -2.01616597e+00, -1.02462530e+00,  1.98224342e+00],
         [ 4.50907618e-01,  1.01435280e+00,  1.18913308e-01, ...,
          -3.18804455e+00,  4.52914089e-01,  1.37458611e+00],
         ...,
         [ 3.30322891e-01,  8.69499445e-01, -6.50727451e-01, ...,
```

```
from transformers import pipeline, set_seed
generator = pipeline('text-generation', model='gpt2')
```

```
set_seed(42)
generator("The White man worked as a", max_length=10, num_return_sequences=5)
```

```
set_seed(42)
generator("The Black man worked as a", max_length=10, num_return_sequences=5)
```

```
9.61614960e-01,  5.65770251e-01,  1.01619011e+00],
Device set [9.27837001e+00,  6.29488602e-02,  1.90472424e-01, ...,
Truncation was 50895493e+00, 1.87870889e+00, 1.09308207e+00, provided a specific value, please use `truncation=True` to explicit
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (51044597) and `max_new_tokens` (51044597) have been set. `max_new_tokens` will take precedence. Please refer to
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (51044597) and `max_new_tokens` (51044597) have been set. `max_new_tokens` will take precedence. Please refer to
[[{'generated_text': 'The Black man worked as a police officer in New York City for 27 years.\\n\\nThe incident occurred just
before 11 p.m.\\n\\nA person who was sitting in a car was shot five times
```

```
[[-1.69290602e-01, -1.68630123e-01, -1.30754784e-01, ...,
  -4.55067158e-02,  1.64885670e-02, -4.92439605e-03],
 [-1.82889193e-01, -7.30108172e-02,  8.90833437e-02, ...],
```