

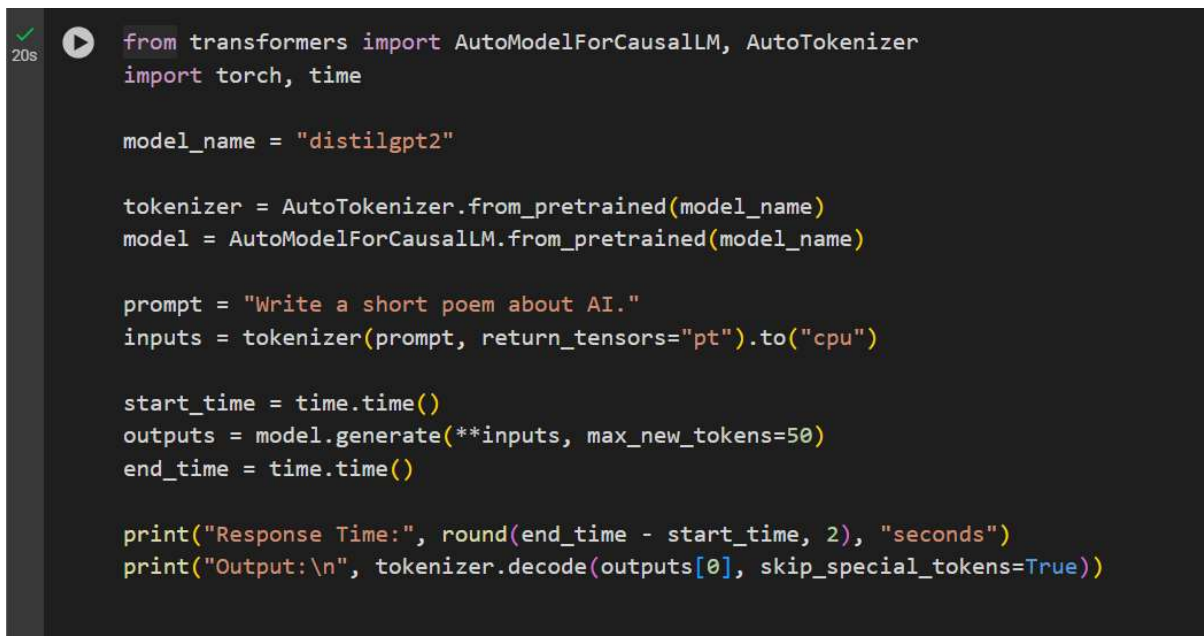
Task1_AI_Poem_LocalLLMWeek3

Assignment 2: Local LLM Installation and Testing

Objective

The objective of this task is to install a local Large Language Model (LLM) in Google Colab and test its functionality by running a simple prompt, measuring response time, and documenting any troubleshooting steps.

Prompt and Output

A screenshot of a Google Colab code cell. On the left, there is a green checkmark icon and a play button icon, with '20s' indicating the execution time. The code cell contains the following Python code:

```
from transformers import AutoModelForCausalLM, AutoTokenizer
import torch, time

model_name = "distilgpt2"

tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name)

prompt = "Write a short poem about AI."
inputs = tokenizer(prompt, return_tensors="pt").to("cpu")

start_time = time.time()
outputs = model.generate(**inputs, max_new_tokens=50)
end_time = time.time()

print("Response Time:", round(end_time - start_time, 2), "seconds")
print("Output:\n", tokenizer.decode(outputs[0], skip_special_tokens=True))
```

