To download the model – 270m specific

- For higher params, use ‘gemma’ library

Steps

1. pip install hf_xet
2. from huggingface_hub import login
   login(token="hf_xxxxxxxxxxx") #with read access

   from huggingface_hub import snapshot_download

   snapshot_download(
     repo_id="google/gemma-3-270m",
     local_dir="gemma-3-270m"
)

   #example output = 'C:\\Users\\Chandan Sagar Rana\\Desktop\\Gemma\\gemma-3-270m'

3. Example code to use it locally
   a. py -3.13 -m pip install --user torch transformers accelerate safetensors

```
from transformers import AutoModelForCausalLM, AutoTokenizer
import torch

path = r"C:\Users\Chandan Sagar Rana\Desktop\Gemma\gemma-3-270m"

# Load tokenizer
tokenizer = AutoTokenizer.from_pretrained(
    path,
    local_files_only=True
)

# Load model
model = AutoModelForCausalLM.from_pretrained(
```

```
    path,
    local_files_only=True,
    device_map="auto",
    torch_dtype=torch.float16 if torch.cuda.is_available() else torch.float32
)

model.eval()

prompt = (
    "Write a full-length, vivid story about the jungle king lion. "
    "The story should have a clear beginning, conflict, and resolution, "
    "with rich descriptions and emotions.\n\nStory:\n"
)

# Tokenize
inputs = tokenizer(
    prompt,
    return_tensors="pt"
)

inputs = {k: v.to(model.device) for k, v in inputs.items()}

# Generate
with torch.no_grad():
    outputs = model.generate(
        **inputs,
        max_new_tokens=500,
        temperature=0.8,            # creativity
        top_p=0.95,                 # nucleus sampling
        do_sample=True,
        repetition_penalty=1.1,     # reduce looping
        eos_token_id=tokenizer.eos_token_id
    )

# Decode
story = tokenizer.decode(outputs[0], skip_special_tokens=True)
print(story)
```

4. Example Output

PS C:\Users\Chandan Sagar Rana\Desktop\Gemma> & "C:/Users/Chandan Sagar Rana/AppData/Local/Programs/Python/Python313/python.exe" "c:/Users/Chandan Sagar Rana/Desktop/Gemma/gemma.py"

`torch_dtype` is deprecated! Use `dtype` instead!

Setting `pad_token_id` to `eos_token_id`:1 for open-end generation.

Write a full-length, vivid story about the jungle king lion. The story should have a clear beginning, conflict, and resolution, with rich descriptions and emotions.

Story:

The king of lions is in the middle of his second month as he's preparing for another big hunting season. He wakes up to find himself on top of a large tree, surrounded by enemies—a pack of giant wild cats. King Lion knows this will be a tough time for him, so he decides that he needs to get away from his usual surroundings. As he tries to climb out of the trees, there are other animals nearby. They don't like this new location, but they do appreciate being able to hide in the shadow of the mountains. After several days, King Lion decides it's high time to go back home. But what if his best friend does something unexpected? Is this the answer to their mutual fears? Or maybe you can help them understand each other better?

Questions:

What would you tell your friends about this scene? What did they see or hear? What was going through their minds as they were watching this?