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

model_id = "key-life/codegen-alpaca-1b"

# Load model & tokenizer
tokenizer = AutoTokenizer.from_pretrained(model_id)
model = AutoModelForCausalLM.from_pretrained(model_id, device_map="auto")

# Example prompt
prompt = "### Instruction:\nWrite a Python function to check if a number is prime.\n\n### Response:\n"
inputs = tokenizer(prompt, return_tensors="pt").to(model.device)

# Generate code
outputs = model.generate(
    **inputs,
    max_new_tokens=128,
    temperature=0.2,        # more deterministic
    top_p=0.9,              # avoids rambling
    do_sample=True,
    eos_token_id=tokenizer.eos_token_id,
    pad_token_id=tokenizer.eos_token_id
)

# Decode
decoded = tokenizer.decode(outputs[0], skip_special_tokens=True)

# Extract only code block
code_block = re.findall(r"```(?:python)?(.*?)```", decoded, re.DOTALL)

if code_block:
    response = code_block[0].strip()
else:
    response = decoded.split("### Response:")[-1].strip()

print(response)
```

```
tokenizer_config.json:  4.14k/?  [00:00<00:00,  226kB/s]
vocab.json:  777k/?  [00:00<00:00,  878kB/s]
merges.txt:  442k/?  [00:00<00:00,  9.16MB/s]
tokenizer.json:  3.48M/?  [00:00<00:00,  318kB/s]
special_tokens_map.json:  100%  906/906  [00:00<00:00,  23.8kB/s]
adapter_config.json:  100%  706/706  [00:00<00:00,  23.8kB/s]
config.json:  100%  1.05k/1.05k  [00:00<00:00,  24.7kB/s]
model.safetensors:  100%  4.55G/4.55G  [00:21<00:00,  182MB/s]
generation_config.json:  100%  111/111  [00:00<00:00,  4.09kB/s]
adapter_model.safetensors:  100%  40.0/40.0  [00:00<00:00,  72.7B/s]
Loading adapter weights from key-life/codegen-alpaca-1b led to missing keys in the model: transformer.h.0.attn.c_attn.lora_A.de
def is_prime(n):
    if n == 2:
        return True
    elif n % 2 == 0:
        return False
    elif n % 3 == 0:
        return False
    else:
        for i in range(5, int(n ** 0.5) + 1, 6):
            if n % i == 0:
                return False
        return True
```

```
Start coding or generate with AI.
```