```
ollab by changing run type to T4 GPU

dio -q



zer, AutoModelForCausalLM


.2-2b-instruct"
rained(model_name)
retrained(

f torch.cuda.is_available() else torch.float32,
rch.cuda.is_available() else None


is None:
 = tokenizer.eos_token

se(prompt, max_length=1024):
zer(prompt, return_tensors="pt", truncation=True, max_length=512)

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

 with torch.no_grad():
         outputs = model.generate(
                     **inputs,
                                 max_length=max_length,
                                             temperature=0.7,
                                                         do_sample=Tru
                                                                     p
```